REPORT DOCUMENTATION PAGE	Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE New Reprint	3. DATES COVERED (From - To) -

4. TITLE AND SUBTITLE
Diffuse Prior Monotonic Likelihood Ratio Test for Evaluation of Fused Image Quality Measures

5a. CONTRACT NUMBER
W911NF-08-1-0449

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Chuanming Wei, Lance M Kaplan, Stephen D Burks, Rick S Blum

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Lehigh University
Office of Research & Sponsored Programs
Lehigh University
Bethlehem, PA 18015 -

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
54261-NS.37

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
This paper introduces a novel method to score how well proposed fused image quality measures (FIQMs) indicate the effectiveness of humans to detect targets in fused imagery. The human detection performance is measured via human perception experiments. A good FIQM should relate to perception results in a monotonic fashion. The method computes a new diffuse prior monotonic likelihood ratio (DPMLR) to facilitate the comparison of the H1 hypothesis that the intrinsic human detection performance is related to the FIQM via a monotonic function against

15. SUBJECT TERMS
Fused image quality measures (FIQM), hypothesistest, image fusion, monotonic correlation (MC).

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Rick Blum
UU	UU	UU	UU		19b. TELEPHONE NUMBER 610-758-3459

Report Title

Diffuse Prior Monotonic Likelihood Ratio Test for Evaluation of Fused Image Quality Measures

ABSTRACT

This paper introduces a novel method to score how well proposed fused image quality measures (FIQMs) indicate the effectiveness of humans to detect targets in fused imagery. The human detection performance is measured via human perception experiments. A good FIQM should relate to perception results in a monotonic fashion. The method computes a new diffuse prior monotonic likelihood ratio (DPMLR) to facilitate the comparison of the H1 hypothesis that the intrinsic human detection performance is related to the FIQM via a monotonic function against the null hypothesis that the detection and image quality relationship is random. The paper discusses many interesting properties of the DPMLR and demonstrates the effectiveness of the DPMLR test via Monte Carlo simulations. Finally, the DPMLR is used to score FIQMs with test cases considering over 35 scenes and various image fusion algorithms.

REPORT DOCUMENTATION PAGE (SF298)
(Continuation Sheet)

Continuation for Block 13

ARO Report Number 54261.37-NS
Diffuse Prior Monotonic Likelihood Ratio Test fo ...

Block 13: Supplementary Note
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

Approved for public release; distribution is unlimited.

Diffuse Prior Monotonic Likelihood Ratio Test for Evaluation of Fused Image Quality Measures

Chuanming Wei, Lance M. Kaplan, *Senior Member, IEEE*, Stephen D. Burks, and Rick S. Blum, *Fellow, IEEE*

***Abstract*—This paper introduces a novel method to score how well proposed fused image quality measures (FIQMs) indicate the effectiveness of humans to detect targets in fused imagery. The human detection performance is measured via human perception experiments. A good FIQM should relate to perception results in a monotonic fashion. The method computes a new diffuse prior monotonic likelihood ratio (DPMLR) to facilitate the comparison of the H_1 hypothesis that the intrinsic human detection performance is related to the FIQM via a monotonic function against the null hypothesis that the detection and image quality relationship is random. The paper discusses many interesting properties of the DPMLR and demonstrates the effectiveness of the DPMLR test via Monte Carlo simulations. Finally, the DPMLR is used to score FIQMs with test cases considering over 35 scenes and various image fusion algorithms.**

***Index Terms*—Fused image quality measures (FIQM), hypothesis test, image fusion, monotonic correlation (MC).**

I. Introduction

In RECENT years, image fusion has been attracting a large amount of attention in a wide variety of applications such as concealed weapon detection [1], remote sensing [2], intelligent robots [3], medical diagnosis [4], and military surveillance [5]. Image fusion refers to generating a combined image in which each pixel is determined from a set of pixels in each of the source images. The fused image should provide an easier view for a human to interpret the scene than any of the source images, thus, improving the performance of the human in accomplishing his/her task. The interested reader is referred to [6, Ch. 1] for a survey of various image fusion algorithms developed in past years.

Measuring the performance of image fusion algorithms is an extremely important task, which has received past study [7]–[22]. The performance of image fusion algorithms is primarily assessed by perceptual evaluation in the form of subjective human tests [13]. Typically in these tests, human observers are asked to view a series of fused images and rate them. Because images are fused for better human interpretation, it is more important to judge fusion methods by how well humans are able to perform interpretation tasks. Examples of human interpretation studies for image fusion evaluations appear in [17], [22]. No matter the goal of the human perception test, these tests are inconvenient, expensive and time consuming.

It is clearly highly desirable to identify an objective performance measure that can accurately predict human perception by determining the quality of the fused image. The objective measure should be a feature that is obtained via an automatic computation employing the fused image and can serve as a surrogate for human perception results. We refer to such a feature as the fused image quality measure (FIQM). If a good FIQM can be devised, then one can compare image fusion algorithms without expensive perception experiments. Furthermore, the measure can be used as a design criteria for an "optimal" image fusion algorithm.

In the literature, three broad classes of FIQMs have been proposed. The first class represents full-reference measures. They require a reference fused image (or the ground truth image) that represents the "ideal" image of the scene. Once the ground truth image is given, one can use existing quality metrics such as the mean square error, the peak signal to noise ratio, or more sophisticated measures such as structure similarity [23] to compare the fused images with the reference. In the image compression application, the uncompressed image represents the ideal, and it has been demonstrated that the structure similarity is a meaningful full-reference measure [23]. For the image fusion application, it is only possible to generate a reference image for some special cases (for instance, the multifocus image fusion [8]). In most cases, one has to resort to other classes of FIQMs that do not require a reference image. The second class of FIQMs represents source comparative measures that utilize partial information about the scene, e.g., the source images that were collected and utilized as input to the image fusion process. This class of FIQMs has recently received a great deal of attention [9]–[12]. These measures quantify the amount of information transferred from the source images to the fused image by considering the sum of correlations between each source image and the fused image. An analysis of this class of FIQMs is provided in [14]. The third class of FIQMs represents no-source comparative measures. These measures attempt to extract the salient fea-

Manuscript received August 18, 2009; revised December 11, 2009 and April 08, 2010; accepted July 06, 2010. Date of publication July 23, 2010; date of current version January 14, 2011. This work was supported by the Air Force Research Laboratory under agreement FA9550-09-1-0576, by the U.S. Army Research Laboratory under Grant W911NF-06-2-0020, by the National Science Foundation under Grant CCF-0829958, by the U.S. Army Research Office under Grant W911NF-08-1-0449, and by a grant from the Commonwealth of Pennsylvania, Department of Community and Economic Development. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhou Wang.

C. Wei and R. S. Blum are with the Electrical and Computer Engineering Department, Lehigh University, Bethlehem, PA 18015 USA (e-mail: chw207@lehigh.edu; rblum@ece.lehigh.edu).

L. M. Kaplan is with the Army Research Laboratory, Adelphi, MD 20783 USA (e-mail: lance.m.kaplan@us.army.mil).

S. D. Burks is with the Development and Engineering Center Night Vision and Electronic Sensors Directorate, Fort Belvoir, VA 22060 USA (e-mail: stephen.burks1@us.army.mil).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2010.2060344

tures, such as the structure, texture, contrast and edge information, directly from the fused image without regard to the source images [17]–[21].

Quantitatively evaluating the image fusion performance is a complicated issue because of the lack of a complete understanding of the human visual system (HVS), and because of the variety of image fusion applications [15]. We expect that the FIQM should be task specific, and the best measure changes from task to task. Given an image fusion application and many kinds of proposed FIQMs, we are interested in which quality measure better describes the performance of the human interpreting the fused imagery.

Ideally, the FIQM for a given image would reveal how well a human can interpret the image for a given task, i.e., it can predict human performance. One can achieve this aim by inventing a measure that linearly fits the human perception performance. In [24], the authors have shown an evidence of the approximately linear fitness between image quality (IQ) measures and the subjective rating of image distortions. However, an image is a projection of a particular scene, and the context in the scene, i.e., the relationship of the objects in the scene, can affect the ability of human to perform a *particular task* (target detection for example). Since the linearity is a stricter requirement than monotonicity for a FIQM and is harder to achieve under various context, we believe that it will be more difficult to guarantee linearity when the IQ is used to predict the ability of a human to interpret the image for a given task. Thus, we focus on the monotonicity criterion in this paper.

By monotonicity we mean that a realistic FIQM can determine the relative ranking of human performance over a series of fused images derived from the same exact source images, which we now refer to as a scene. For a given scene, as FIQM increases over a series of fused images, human performance over these images should also increase. If the human performance is consistently decreasing, the measure is still good as it can be trivially transformed into a proper FIQM via a reciprocal operation. Thus, a potential FIQM should be judged by how well a monotonic function (ascending or descending) explains the relation between the FIQM and human performance over a variety of fused imagery representing the same scene. In addition, the nature of the monotonic relationship (ascending or descending) should be consistent from scene to scene. Overall, a statistic that quantifies how well different FIQMs are consistent with actual human performance is necessary.

This paper focuses on scoring FIQMs for the case of the detection task. Performance is measured by the probability that a human observer can correctly detect certain objects in the fused image. The human perception experiments measure the number of observers that are able to correctly detect ground truthed targets as the human performance. This performance metric can be reasonably modeled by a binomial distribution. This paper introduces a new monotonic statistic for the object detection task that is applicable when the underlying perception results are derived from a small number of human observers. To handle a small number of observers, this statistic does not make Gaussian assumptions about the performance measurements.

Previous work does exist to objectively score the effectiveness of FIQMs. In [16], Pearson (or linear) correlation and root mean squared error (RMSE) are used to score potential FIQMs. The Pearson correlation quantifies how well a straight line fits the mapping between the input and output sequences. Unfortunately, when the relationship between the quality measure and the human performance is nonlinear, the value of Pearson correlation can be small despite the fact that the sequences are still monotonically related. In essence, a proper statistic needs to determine if the ordering of a quality measure preserves the ordering of the corresponding human performance measures.

The Spearman and Kendall correlations [25], [26] are common statistics to quantify how well the output sequence is ordered. In fact, the Spearman correlation has been used to evaluate the quality measures for video streams [27]. Both quantities are invariant to monotonic transformations of both the input and output sequences because monotonic transformations preserve the rank order of the sequences. For evaluation of the utility of FIQMs, a miss-ordering of human performance values that are nearly identical should not lower the correlation value too much. Because only ranks and not actual values are considered, the reduction in correlation score due to these miss-orderings can be identical or even greater than that of miss-orderings of widely varying human performance values. Clearly, measurement noise can greatly impact the correlation scores. Therefore, these rank-order correlations are not appropriate for seeking out good FIQMs.

In [23], [27], a nonlinear regression fit to a logistic function followed by linear correlation is used to compare various FIQMs in order to accommodate the nonlinear, but monotonic, relationships. Recently, the monotonic correlation (MC), which uses isotonic regression followed by linear correlation has been proposed in [17]. As demonstrated in [17], the MC is more flexible than linear correlation or the logistic analysis in [23], [27]. Like linear and logistic correlation, the MC assumes that the perception error is Gaussian, which is inappropriate for the detection task when the number of observers is small.

To our knowledge, this paper represents the first attempt to score the effectiveness of FIQMs for the detection task in light of practical measurements from human perception experiments. To this end, the paper develops a novel statistic to test whether or not a monotonic relationship exists between the proposed FIQM and the human performance. The monotonic statistic is general and can be applied to other applications when one may need to test for a monotonic relationship. A preliminary version of this work has appeared in [28].

The paper is organized as follows. Section II presents the perception model and introduces the new monotonic statistic. Section III demonstrates the effectiveness of the new statistic via Monte Carlo simulations. The statistic is used to score potential FIQMs against actual perception results for fused images in Section IV. Finally, Section V provides some concluding remarks.

II. Statistical Monotonic Analysis

The paper focuses on the detection task and measures the performance of image fusion algorithms by the probability that a human observer can correctly detect certain objects in the fused image. This section develops the test statistic that compares the hypothesis that the relationship between human detection performance and FIQM values are monotonic to the hypothesis that the relationship is random. The statistic is based upon the model that each image exhibits a ground truth quality score, which is the probability that any human can detect the object in it. Section II-A derives the likelihoods for each hypothesis conditioned on these ground truth quality scores. Then, Section II-B uses an uninformative prior for the ground truth quality scores to define the likelihood ratio so that it is computationally feasible as demonstrated in Section II-C. Finally, Section II-D presents properties of the test statistic.

A. Data Models

A scene is a realization of F source images, and we consider N different fusion algorithms. The existence (or lack) of a monotonic relationship between measured human performance and computed FIQMs can be inferred over S scenes. To this end, this subsection provides the data models that enable this inference.

For a given scene, let the $N\times 1$ vector $\tilde{\mathbf{p}} = (\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_N)^T$ denote the actual performance for all fusion methods, where $\tilde{p}_i$ is the object detection probability, i.e., the ground truth quality score, associated with the image obtained from the ith fusion algorithm. Let a given FIQM evaluated over N fusion algorithms be denoted as a $N\times 1$ vector $\tilde{\mathbf{x}} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_N)^T$. The computed value $\tilde{x}_i$ is a deterministic function of the image obtained from the ith fusion algorithm and the F source images. The proposed monotonic hypothesis test evaluates how well a FIQM monotonically relates to human object detection performance. Under the monotonic hypothesis, there is a monotonic function that maps the measure value $\tilde{x}_i$ associated with the ith fusion algorithm to the detection probability $\tilde{p}_i$, i.e.,

$$\tilde{p}_i = g(\tilde{x}_i) \tag{1}$$

where $g(x)$ is a monotonic increasing or decreasing function of x. Let $\mathbf{p}$ and $\mathbf{x}$ denote a reordering of $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{x}}$ such that the measure values are in ascending order. i.e., $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_N$. Note that $\mathbf{p} = P_k\tilde{\mathbf{p}}$ and $\mathbf{x} = P_k\tilde{\mathbf{x}}$ where P_k is one of a possible $N!$ permutation matrices. This paper uses the convention that P_1 is the identity matrix and $P_{N!}$ reverses the original ordering, i.e., the anti-diagonal matrix of ones. Now, we consider two alternative H_1 hypotheses: $H_\uparrow$ for ascending p_i's and $H_\downarrow$ for descending p_i's. On the other hand, the null hypothesis is that over the ensemble of possible fused imagery, the $\tilde{x}_i$'s are *i.i.d.* samples. Thus, the p_i's are in random order where the probability of any permutation of the order is equal. In other words, P_k is the permutation matrix that orders the p_i's for the H_1 hypotheses, and P_k is randomly chosen via a uniform distribution over the $N!$ possible permutation matrices under the null (H_0) hypothesis. Namely, the conditional probability mass functions (pmfs) of the permutations conditioned on $\tilde{\mathbf{p}}$ and the hypotheses for $k = 1, \ldots, N!$ are

$$\begin{aligned} f_\pi(P_k \,|\, \tilde{\mathbf{p}}, H_\uparrow) &= \begin{cases} 1, & \text{if } P_k\tilde{\mathbf{p}} \in \mathcal{P}_\uparrow \\ 0, & \text{otherwise,} \end{cases} \\ f_\pi(P_k \,|\, \tilde{\mathbf{p}}, H_\downarrow) &= \begin{cases} 1, & \text{if } P_k\tilde{\mathbf{p}} \in \mathcal{P}_\downarrow \\ 0, & \text{otherwise} \end{cases} \\ f_\pi(P_k \,|\, \tilde{\mathbf{p}}, H_0) &= \frac{1}{N!} \end{aligned} \tag{2}$$

where

$$\begin{aligned} \mathcal{P}_\uparrow &= \{\mathbf{p} : 0 \leqslant p_1 \leqslant \ldots \leqslant p_N \leqslant 1\} \\ \mathcal{P}_\downarrow &= \{\mathbf{p} : 1 \geqslant p_1 \geqslant \ldots \geqslant p_N \geqslant 0\}. \end{aligned} \tag{3}$$

For this discusion, it is also convenient to define $\mathcal{P}_0$ as the set of all possibe $\mathbf{p}$'s, i.e.,

$$\mathcal{P}_0 = \{\mathbf{p} : 0 \leqslant p_1, \ldots, p_N \leqslant 1\}. \tag{4}$$

If $\mathbf{p} = P_k\tilde{\mathbf{p}}$ is observed, then the likelihoods of the hypotheses, i.e., $l(H_i \,|\, \mathbf{p}) = f(P_k \,|\, \tilde{\mathbf{p}}, H_i)$ for $i \in \{\uparrow, \downarrow, 0\}$ demonstrate that if $\mathbf{p}$ is not in ascending (or descending) order, then the ascending (or descending) likelihood (and likelihood ratio) is zero, and the $H_\uparrow$ (or $H_\downarrow$) hypothesis must be incorrect. Also, if $\mathbf{p}$ happens to be in ascending (or descending) order, then either the $H_\uparrow$ (or $H_\downarrow$) hypothesis is true or the ordering of $\mathbf{p}$ is due to random luck under the null hypothesis, which occurs with a probability of $1/N!$. Thus, for $\mathbf{p} \in \mathcal{P}_\uparrow$ (or $\mathbf{p} \in \mathcal{P}_\downarrow$), the likelihood ratio is not infinite, i.e., a sure monotonic relationship. Rather, it is $N!$ due to the fact that the random $\mathbf{x}$ can order $\mathbf{p}$ by chance.

Unfortunately, the value of $\tilde{\mathbf{p}}$ (or $\mathbf{p}$) is unobservable. It can only be inferred via perception experiments that measure $\mathbf{y} = (y_1, y_2, \ldots, y_N)^T$ where y_i is the number of observers that correctly detect the targets in the image obtained from the ith fusion algorithm.[1] We use o_i to represent the number of observers that participate in the detection experiment for the image formed by the ith fusion image. Under the assumption that all human are equally capable, it is reasonable to model $\mathbf{y}$ as a random vector whose elements are statistically independent where y_i is drawn from a binomial distribution with parameters o_i and p_i so that the pmf of $\mathbf{y}$ conditioned on $\mathbf{o}$ and $\mathbf{p}$ is

$$\mathbf{y} \sim f_y(\mathbf{y} \,|\, \mathbf{o}, \mathbf{p}) = \prod_{i=1}^{N} \binom{o_i}{y_i} p_i^{y_i} (1 - p_i)^{o_i - y_i}. \tag{5}$$

Here we represent the o_i's in an $N \times 1$ vector $\mathbf{o}$ for notational convenience. Since $\mathbf{p} = P_k\tilde{\mathbf{p}}$, one can define $f_y(\mathbf{y} \,|\, \mathbf{o}, P_k, \tilde{\mathbf{p}}) = f_y(\mathbf{y} \,|\, \mathbf{o}, \mathbf{p})$.

The joint pmf of the observations $\mathbf{y}$ and the permutations P_k can be written as

$$f_{y\pi}(\mathbf{y}, P_k \,|\, \mathbf{o}, \tilde{\mathbf{p}}, H_i) = f(\mathbf{y} \,|\, P_k, \mathbf{o}, \tilde{\mathbf{p}}, H_i) f(P_k \,|\, \mathbf{o}, \tilde{\mathbf{p}}, H_i). \tag{6}$$

Because $\mathbf{y}$ conditioned on $\mathbf{o}$ and $\tilde{\mathbf{p}}$ is independent of H_i, $f(\mathbf{y} \,|\, P_k, \mathbf{o}, \tilde{\mathbf{p}}, H_i) = f_y(\mathbf{y} \,|\, P_k, \mathbf{o}, \tilde{\mathbf{p}})$ for all H_i's. Furthermore, $f(P_k \,|\, \mathbf{o}, \tilde{\mathbf{p}}, H_i) = f_\pi(P_k \,|\, \tilde{\mathbf{p}}, H_i)$ because P_k does not depend upon $\mathbf{o}$. Thus, $f_{y\pi}(\mathbf{y}, P_k \,|\, \mathbf{o}, \tilde{\mathbf{p}}, H_i)$ is obtained by the multiplication of (2) and (5) so that

$$\begin{aligned} f_{y\pi}(\mathbf{y}, P_k \,|\, \mathbf{o}, \tilde{\mathbf{p}}, H_\uparrow) &= \begin{cases} f_y(\mathbf{y} \,|\, \mathbf{o}, P_k, \tilde{\mathbf{p}}), & \text{if } P_k\tilde{\mathbf{p}} \in \mathcal{P}_\uparrow \\ 0, & \text{otherwise} \end{cases} \\ f_{y\pi}(\mathbf{y}, P_k \,|\, \mathbf{o}, \tilde{\mathbf{p}}, H_\downarrow) &= \begin{cases} f_y(\mathbf{y} \,|\, \mathbf{o}, P_k, \tilde{\mathbf{p}}), & \text{if } P_k\tilde{\mathbf{p}} \in \mathcal{P}_\downarrow \\ 0, & \text{otherwise} \end{cases} \\ f_{y\pi}(\mathbf{y}, P_k \,|\, \mathbf{o}, \tilde{\mathbf{p}}, H_0) &= \frac{1}{N!} f_y(\mathbf{y} \,|\, \mathbf{o}, P_k, \tilde{\mathbf{p}}). \end{aligned} \tag{7}$$

[1] For variables that do not use the tilde, the indices for the images are such that $\mathbf{x}_i$'s are in ascending order.

Then, a hypothesis test to distinguish $H_\uparrow$ or $H_\downarrow$ from H_0 using the observed values can be derived from the likelihoods $l(H_i \,|\, \mathbf{y}, \mathbf{o}, P_k, \tilde{\mathbf{p}}) = f_{y\pi}(\mathbf{y}, P_k \,|\, \mathbf{o}, \tilde{\mathbf{p}}, H_i)$. Because $\mathbf{p} = P_k \tilde{\mathbf{p}}$ is not observed, the hypothesis test is a composite test. It is unclear whether a uniformly most powerful (UMP) test exists. A common test to apply is the generalized likelihood ratio test (GLRT). This requires one to compute the maximum likelihood (ML) estimates $\hat{\mathbf{p}}_\uparrow$, $\hat{\mathbf{p}}_\downarrow$, $\hat{\mathbf{p}}_0$ for the $H_\uparrow$, $H_\downarrow$, and H_0 hypotheses, respectively. For the two H_1 hypotheses, the ML estimates can be obtained by the $O(N)$ pool adjacent violators algorithm [17], [29], [30]. For the null hypothesis, $\hat{p}_i = (y_i/o_i)$. The GLRT has the property that for any ascending (or descending) $\mathbf{y}$, the ascending (or descending) generalized likelihood ratio (GLR) is $N!$. However, if the y_i's are close in values, the ordering is more likely to be due to luck than when the y_i's are well spread. However, the GLRT is unable to make this distinction between different ordered $\mathbf{y}$'s. A different approach that accounts for the relative spread of the observations values is needed.

B. *Diffuse Prior Monotonic Likelihood Ratio Test*

A given scene is a realization from the ensemble of possible source images. Therefore, it is reasonable to model the detection probabilities as being drawn from a random distribution, i.e., $\tilde{\mathbf{p}} \sim f_{\tilde{p}}(\tilde{\mathbf{p}})$. The *diffuse prior monotonic likelihood ratio test* (DPMLRT) assumes that for a given scene, $\tilde{\mathbf{p}}$ is a realization of an uninformative (or diffuse) prior distribution, i.e., the elements $\tilde{p}_i$ are i.i.d. uniform $[0, 1)$ so that $f_{\tilde{p}}(\tilde{\mathbf{p}}) = 1$. The uniform distribution models the fact that the imagery are collected in various conditions where the ability to detect the objects can be easy, hard, or somewhere in between. The independence between fusion methods is a simplifying assumption that leads to a computationally feasible test. Because the prior on $\tilde{\mathbf{p}}$ is independent of the hypothesis H_i and $\mathbf{o}$, we have $f(\tilde{\mathbf{p}} \,|\, \mathbf{o}, H_i) = f_{\tilde{p}}(\tilde{\mathbf{p}}) = 1$. Then, $\tilde{\mathbf{p}}$ is marginalized so that the expected likelihood for the ith hypothesis is

$$\begin{aligned} \tilde{l}(H_i \,|\, \mathbf{y}, \mathbf{o}, P_k) &= \int_{\mathcal{P}_0} f_{y\pi}(\mathbf{y}, P_k \,|\, \mathbf{o}, \tilde{\mathbf{p}}, H_i) f(\tilde{\mathbf{p}} \,|\, \mathbf{o}, H_i) d\tilde{\mathbf{p}} \\ &= \int_{\mathcal{P}_0} f_{y\pi}(\mathbf{y}, P_k \,|\, \mathbf{o}, \tilde{\mathbf{p}}, H_i) d\tilde{\mathbf{p}}. \end{aligned} \tag{8}$$

Now the expected likelihoods do not depend upon any unobservable parameters. The integral in (8) can be simplified by noting that the integrand is given by (7) and using the change of variable $\tilde{\mathbf{p}} \mapsto P_k^{-1} \mathbf{p}$. Then, it is easy to see that

$$\begin{aligned} \tilde{l}(H_\uparrow \,|\, \mathbf{y}, \mathbf{o}, P_k) &= \int_{\mathcal{P}_\uparrow} f_y(\mathbf{y} \,|\, \mathbf{o}, \mathbf{p}) d\mathbf{p} \\ \tilde{l}(H_\downarrow \,|\, \mathbf{y}, \mathbf{o}, P_k) &= \int_{\mathcal{P}_\downarrow} f_y(\mathbf{y} \,|\, \mathbf{o}, \mathbf{p}) d\mathbf{p} \\ \tilde{l}(H_0 \,|\, \mathbf{y}, \mathbf{o}, P_k) &= \frac{1}{N!} \int_{\mathcal{P}_0} f_y(\mathbf{y} \,|\, \mathbf{o}, \mathbf{p}) d\mathbf{p}. \end{aligned} \tag{9}$$

Now, the tests to distinguish the H_1 hypotheses from the null hypothesis are simple hypothesis tests, and the likelihood ratio test (LRT) is the most powerful test. Namely, given that for each scene $\tilde{\mathbf{p}}$ is drawn from the uninformative prior, then the following LRTs are optimal in the Neyman-Pearson sense [31] for distinguishing the monotonically ascending or descending hypothesis from the null hypothesis[2]

$$\begin{aligned} \lambda_N^\uparrow(\mathbf{y}, \mathbf{o}) &= \frac{N! \int_{\mathcal{P}_\uparrow} f(\mathbf{y} \,|\, \mathbf{o}, \mathbf{p}) d\mathbf{p}}{\int_{\mathcal{P}_0} f(\mathbf{y} \,|\, \mathbf{o}, \mathbf{p}) d\mathbf{p}} \\ \lambda_N^\downarrow(\mathbf{y}, \mathbf{o}) &= \frac{N! \int_{\mathcal{P}_\downarrow} f(\mathbf{y} \,|\, \mathbf{o}, \mathbf{p}) d\mathbf{p}}{\int_{\mathcal{P}_0} f(\mathbf{y} \,|\, \mathbf{o}, \mathbf{p}) d\mathbf{p}}. \end{aligned} \tag{10}$$

We refer to $\lambda_N^\uparrow$ and $\lambda_N^\downarrow$ as the ascending and descending diffuse prior monotonic likelihood ratio (DPMLR), respectively.

For multiple scenes, the nature of the monotonicity (ascending or descending) should be consistent from scene to scene. Therefore, one must consider the cumulative likelihoods for the ascending, descending, and null hypotheses. Since we assume that the $\mathbf{y}$'s and $\mathbf{p}$'s are statistically independent from scene to scene, the likelihoods for each hypothesis accumulate via the product operation. The cumulative likelihood ratios are then proportional to the geometric mean of the likelihood ratios for each scene. The geometric mean provides a convenient way to normalize the score against the number of scenes. The overall likelihood ratio for the monotonic relationship over S scenes is formally defined as

$$\Lambda_N = \left(\max \left\{ \prod_{s=1}^{S} \lambda_N^\uparrow(\mathbf{y}_s, \mathbf{o}_s), \prod_{s=1}^{S} \lambda_N^\downarrow(\mathbf{y}_s, \mathbf{o}_s) \right\} \right)^{1/S} \tag{11}$$

where $\mathbf{y}_s$ and $\mathbf{o}_s$ are the number of correct detections and observers for the sth scene, respectively. Note that Λ_N is agnostic to the nature of the monotonicity. Unless it is required, the scene index is implicit for the sake of notational brevity. We refer to Λ_N as the composite DPMLR. When $\Lambda_N > 1$ the evidence in support of the monotonic hypothesis is greater than that of the null hypothesis where the FIQM behaves as noise with respect to human performance. As Λ_N increases, so does the evidence that the FIQM under test is actually a good measure. The DPMLRT is simply accepting the monotonic hypothesis if the DPMLR exceeds a given threshold value. Usually, the threshold is greater than one.

C. *Recursive Computation*

To our knowledge, a closed form expression for (10) does not exist, and numerical integration quickly becomes infeasible as N increases. Fortunately, it is possible to calculate the diffuse likelihood ratios numerically. However, due to the multivariable integration involved in the expression, the calculation requires large computational cost, especially when N and the o_i's are large. This subsection provides a recursion to calculate these diffuse likelihood ratios.

The diffuse likelihood for H_0 can be simply expressed as

$$\tilde{l}(H_0 \,|\, \mathbf{y}, \mathbf{o}) = \prod_{i=1}^{N} \binom{o_i}{y_i} \beta(y_i + 1, o_i - y_i + 1) \tag{12}$$

[2] For notational convenience, the dependency of λ to the ordering P_k is left implicit since λ is actually invariant to P_k except in how it orders $\mathbf{y}$.

where

$$\beta(a,b)=\int_0^1 z^{a-1}(1-z)^{b-1}\,dz \tag{13}$$

is the Beta function.

Substituting equations (5), (8) and (12) into (10), the ascending diffuse likelihood ratio can be expressed as

$$\lambda_N^{\uparrow}(\mathbf{y},\mathbf{o})=\frac{N!\int_{\mathcal{P}_\uparrow} h(p_N;y_N,o_N)\ldots h(p_1;y_1,o_1)d\mathbf{p}}{\prod_{i=1}^N \beta(y_i+1,o_i-y_i+1)} \tag{14}$$

where

$$h(p;y,o)=p^y(1-p)^{o-y}. \tag{15}$$

By considering the power series expansion of the regularized incomplete Beta function, the calculation of $\lambda_N^{\uparrow}(\mathbf{y},\mathbf{o})$ can be simplified in a recursive way. Specifically, the regularized incomplete Beta function is defined as

$$I(y;a,b)=\frac{\int_0^y z^{a-1}(1-z)^{b-1}\,dz}{\beta(a,b)} \tag{16}$$

and the power series expansion for $I(y;a,b)$ is

$$I(y;a,b)=\frac{1}{a+b}\times\sum_{j=a}^{a+b-1}\frac{1}{\beta(j+1,a+b-j)}y^j(1-y)^{a+b-1-j}. \tag{17}$$

Then, (14) can be written as (18), shown at the bottom of the page. Now substituting (17) into (18), we obtain (19), shown at the bottom of the page.

Also from (3), one can see that $\mathcal{P}_\uparrow$ and $\mathcal{P}_0$ are the same when $N=1$. Therefore, by definition, we have

$$\lambda_1^{\uparrow}(y_1,o_1)=1 \tag{20}$$

and the ascending diffuse likelihood ratio can be computed numerically via the recursion defined in (19) and (20). A similar recursion can compute the descending diffuse likelihood ratio. Alternatively, one can use the symmetry property (see Property 2 in the next subsection) to derive $\lambda_N^{\downarrow}$ from the computation of $\lambda_N^{\uparrow}$.

D. Properties

The diffuse likelihood ratios demonstrate a number of interesting properties than can easily be proven. Some of these properties are for the general case where the number of observers can vary over the different fused images. Other properties are for the case that the number of observers is constant, i.e., $o_i=o$. This more specific case that $\mathbf{o}=o\mathbf{1}$ is common for perception experiments where one would expect the evaluation of the fused imagery over the same number of observers. In addition to these provable properties, we have discovered other interesting attributes for the DPMLR by exhaustively computing the DPMLRs for all $(o+1)^N$ values of $\mathbf{y}$ for manageable, i.e., small, values of o and N. These attributes make sense based upon the intuition of how the DPMLRT should behave; we speculate that these attributes are preserved for larger values of o and N; and we are willing to go out on a limb by disseminating them as *conjectures* in this subsection. We hope that proofs will be discovered in the future so that the conjectures can become properties.

This section first presents the properties that are valid for general values of $\mathbf{o}$.

Property 1: $\lambda_N^{\uparrow}(\mathbf{y},\mathbf{o}),\lambda_N^{\downarrow}(\mathbf{y},\mathbf{o}),\Lambda_N\in(0,N!)$.

The proof of this property can be found in Appendix A. The property bounds the possible values of the diffuse likelihood

$$\frac{N!\int_0^1\cdots\int_0^{p_3} h(p_N;y_N,o_N)\ldots h(p_2;y_2,o_2)\left(\int_0^{p_2} h(p_1;y_1,o_1)dp_1\right)dp_2\ldots dp_N}{\beta(y_1+1,o_1-y_1+1)\prod_{i=2}^N\beta(y_i+1,o_i-y_i+1)} = \frac{N!\int_0^1\cdots\int_0^{p_3} h(p_N;y_N,o_N)\ldots h(p_2;y_2,o_2)I(p_2;y_1+1,o_1-y_1+1)dp_2\ldots dp_N}{\prod_{i=2}^N\beta(y_i+1,o_i-y_i+1)} \tag{18}$$

$$\begin{aligned}\lambda_N^{\uparrow}(\mathbf{y},\mathbf{o})&=\frac{N!}{o_1+2}\sum_{j=y_1+1}^{o_1+1}\frac{\beta(j+y_2+1,o_1+o_2+2-y_2-j)}{\beta(j+1,o_1+2-j)\beta(y_2+1,o_2-y_2+1)}\\&\quad\times\frac{\int_0^1\cdots\int_0^{p_3} h(p_N;y_N,o_N)\ldots h(p_2;j+y_2,o_1+o_2+1)\,dp_2\ldots dp_N}{\prod_{i=3}^N\beta(y_i+1,o_i-y_i+1)\beta(j+y_2+1,o_1+o_2+2-y_2-j)}\\&=\frac{N!}{o_1+2}\sum_{j=y_1+1}^{o_1+1}\frac{\beta(j+y_2+1,o_1+o_2+2-y_2-j)}{\beta(j+1,o_1+2-j)\beta(y_2+1,o_2-y_2+1)}\\&\quad\times\lambda_{N-1}^{\uparrow}\left([j+y_2,y_3,\ldots,y_N]',[o_1+o_2+1,o_3,\ldots,o_N]'\right)\end{aligned} \tag{19}$$

ratios. As the number of objects N to consider increases, the upper bound for the likelihood ratios grows fast. For a given value of N and o, the bounds of zero and $N!$ are loose since the set of all possible values of $\mathbf{y}$ is finite. However, as demonstrated later in this subsection, as the number of observers increases, one can find a $\mathbf{y}$ that corresponds to a likelihood ratio value that is arbitrarily close to either bound. In other words, as the number of observers increases and the y_i's have sufficient spread, the likelihood ratio becomes as if $\mathbf{p}$ is observable (see Section II-A).

Property 2: $\lambda_N^{\downarrow}(\mathbf{y}, \mathbf{o}) = \lambda_N^{\uparrow}(P_{N!}\mathbf{y}, P_{N!}\mathbf{o}) = \lambda_N^{\uparrow}(\mathbf{o} - \mathbf{y}, \mathbf{o})$.

Proof: The first equality is the result of a simple change of variables $\mathbf{p} \mapsto P_{N!}\mathbf{p}$ in (14). Likewise, the second equality is the result of the change of variables $p_i \mapsto 1 - p_i$ for $i = 1, \ldots, N$ in (14) followed by a reversal of the order of integration. ■

This property demonstrates a symmetry between $\lambda_N^{\uparrow}$ and $\lambda_N^{\downarrow}$. The symmetry provides a convenient way to derive the descending likelihood ratio via the computation of the ascending likelihood ratio and vice versa.

The first two properties are valid for a variable amount of observers per a fused image. The final set of properties are specific for the case that a constant number of observers o are utilized for the N fused images, i.e., $\mathbf{o} = o\mathbf{1}$.

Property 3: If $y_1 = y_2 = \cdots = y_N$, then $\lambda_N^{\downarrow}(\mathbf{y}, o\mathbf{1}) = \lambda_N^{\uparrow}(\mathbf{y}, o\mathbf{1}) = 1$.

The proof of this property is given in Appendix B. The property states that when all observations are equal, one cannot distinguish between the ascending, descending, and null hypotheses because all orderings of the observations are indistinguishable. Clearly, when all observations are the same, it is an ill-posed problem to determine whether or not the FIQMs are ordering the fused imagery in any special manner.

Property 4: If the y_i's are in ascending order and they are not constant then $\lambda_N^{\uparrow}(P_{N!}\mathbf{y}, P_{N!}\mathbf{o}) < \lambda_N^{\uparrow}(P_k\mathbf{y}, P_k\mathbf{o}) < \lambda_N^{\uparrow}(\mathbf{y}, \mathbf{o})$ for $1 < k < N!$. Likewise, if the y_i's are in descending order and they are not constant then $\lambda_N^{\downarrow}(P_{N!}\mathbf{y}, P_{N!}\mathbf{o}) < \lambda_N^{\downarrow}(P_k\mathbf{y}, P_k\mathbf{o}) < \lambda_N^{\downarrow}(\mathbf{y}, \mathbf{o})$ for $1 < k < N!$.

Property 5: If the y_i's are in ascending order and they are not constant, then $\lambda_N^{\uparrow}(\mathbf{y}, \mathbf{o}) > 1$ and $\lambda_N^{\downarrow}(\mathbf{y}, \mathbf{o}) < 1$. Likewise, if the y_i's are in descending order and they are not constant, then $\lambda_N^{\downarrow}(\mathbf{y}, \mathbf{o}) > 1$ and $\lambda_N^{\uparrow}(\mathbf{y}, \mathbf{o}) < 1$.

The proof of these two properties is provided in Appendix C. Property 4 states that if the observations demonstrate a perfect monotonic ascending relationship with the FIQM, then the ascending likelihood ratio is larger than that for any other ordering of the observations. Furthermore, the descending order of observations demonstrates the lowest ascending likelihood ratio of all possible orderings. Since it can be shown that the average likelihood ratio over all possible orderings of the observations is one, Property 5 is a corollary of Property 4. The property states that as long as the human performance $\mathbf{y}$ is increasing in concert with $\mathbf{x}$, the diffuse likelihood ratio will always favor the ascending $H_{\uparrow}$ and disfavor the descending $H_{\downarrow}$ hypotheses over the null hypothesis H_0. Similarly, as long as the human performance $\mathbf{y}$ is decreasing in concert with $\mathbf{x}$, the diffuse likelihood ratio will always favor the descending $H_{\downarrow}$ and disfavor the ascending $H_{\uparrow}$ hypotheses over the null hypothesis H_0. Clearly, these two properties are both intuitively appealing.

Conjecture 1: The product $\lambda_N^{\uparrow}(\mathbf{y}, o\mathbf{1}) \cdot \lambda_N^{\downarrow}(\mathbf{y}, o\mathbf{1}) \leqslant 1$ where equality occurs if and only if $\lambda_N^{\uparrow}(\mathbf{y}, o\mathbf{1}) = \lambda_N^{\downarrow}(\mathbf{y}, o\mathbf{1}) = 1$.

As stated earlier, this conjecture is the result of searching through an exhaustive list of $(o+1)^N$ monotonic likelihood ratio values for manageable values of o and N. This conjecture states that the ascending and descending hypotheses can never both be favored over the null hypothesis. In other words, $\lambda_N^{\uparrow} > 1$ implies $\lambda_N^{\downarrow} < 1$, and $\lambda_N^{\downarrow} > 1$ implies $\lambda_N^{\uparrow} < 1$. However, the converse is not true. It is possible that for a given $\mathbf{y}$ both $\lambda_N^{\uparrow}$ and $\lambda_N^{\downarrow}$ can be less than one. As a simple example, consider $\mathbf{y} = [0\ 2\ 0]$ for $o = 2$. Because of the symmetry property, $\lambda_N^{\uparrow} = \lambda_N^{\downarrow}$. At best, a symmetric $\mathbf{y}$ can have a monotonic likelihood ratio of one when all the y_i's are constant. Otherwise, the symmetric $\mathbf{y}$ is neither ascending or descending and should not provide evidence to support $H_{\uparrow}$ or $H_{\downarrow}$ over H_0. For this case, the ascending, descending, and composite DPMLRs are all 0.2286.

Conjecture 2: $\lambda_N^{\uparrow}(\mathbf{y}, o\mathbf{1}) = 1$ (or $\lambda_N^{\downarrow}(\mathbf{y}, o\mathbf{1}) = 1$) if and only if the y_i's are constant.

This conjecture states that the only way for the ascending (or descending) hypothesis to be indistinguishable from the null hypothesis is for all the observations y_i to be the same. Furthermore, if the ascending hypothesis cannot be distinguished from the null hypothesis then the same is true for the descending hypothesis.

Conjecture 3: For a given N, the bounds in Property 1 are tight in the sense that one can identify a value of o and corresponding $\mathbf{y}$ whose monotonic likelihood ratio is arbitrarily close to either the lower bound of zero or the upper bound of $N!$.

Inspection of the exhaustive list of monotonic likelihood ratios of possible $\mathbf{y}$'s for small values of N and o has revealed that

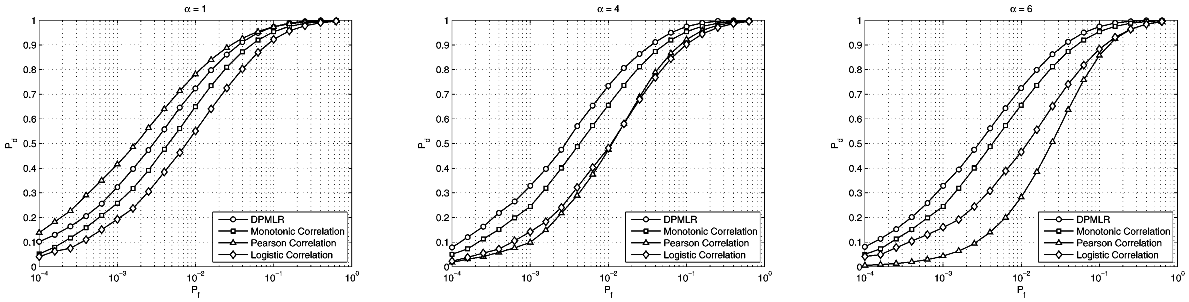
$$\bar{y}_i = \left\lfloor \frac{i-1}{N-1} o \right\rfloor \quad \text{and} \quad \underline{y}_i = \begin{cases} o, & i < N/2 \\ 0, & i \geqslant N/2 \end{cases} \tag{21}$$

achieve close to the maximum and minimum values of $\lambda_N^{\uparrow}$, respectively, for a given value of N and o. A different rounding function in (21) may lead to a higher $\lambda_N^{\uparrow}$. Intuitively, as the values of the y_i's spread apart, the discriminability between the hypotheses improves. If the observations use the entire dynamic range of o and they increase linearly with respect to the rank order, then it makes sense that $\lambda_N^{\uparrow}$ is as large as possible. Since maximizing $\lambda_N^{\uparrow}$ also maximizes Λ_N due to (11) and the symmetry property, $\bar{\mathbf{y}}$ also achieves close to the maximum of Λ_N. For a small $\lambda_N^{\uparrow}$, the y_i's should be decreasing and $\underline{\mathbf{y}}$ has the maximum drop possible. While $\underline{\mathbf{y}}$ leads to a small $\lambda_N^{\uparrow}$, its corresponding Λ_N value is greater than one because it is monotonically descending [see (11)]. The observation sequence

$$\breve{y}_i = (1 - (-1)^i)o/2 \tag{22}$$

achieves close to the minimum value of Λ_N for a given value of N and o. It is neither increasing nor decreasing and utilizes the dynamic range of o. Table I demonstrates how these sequence are converging to the lower and upper bounds for $\lambda_N^{\uparrow}$ and Λ_N as o increases for a given N. The symmetry properties can be used to show similar results for $\lambda_N^{\downarrow}$.

In summary, the evidence to accept the H_1 hypothesis $(\text{DPMLR} > 1)$ or the null hypothesis $(\text{DPMLR} < 1)$ increases as the number of observers increases because the spread



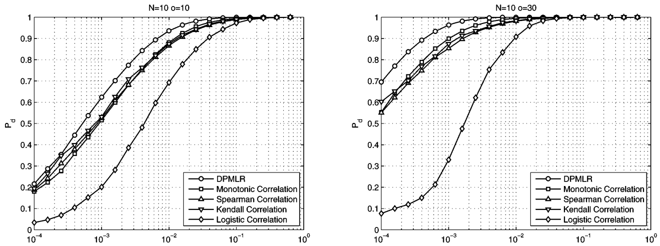
(a) (b) (c)

Fig. 1. ROC curves for DPMLR, MC, Pearson correlation, and logistic correlation tests. (a) $\alpha = 1$. (b) $\alpha = 4$. (c) $\alpha = 6$.

TABLE I
VALUES OF $\bar{\mathbf{y}}$, $\underline{\mathbf{y}}$, AND $\breve{\mathbf{y}}$ SHOW THAT $\lambda_N^{\uparrow}$ AND Λ_N CAN APPROACH THEIR BOUNDS OF ZERO AND $N!$ AS THE NUMBER OF OBSERVERS o INCREASES

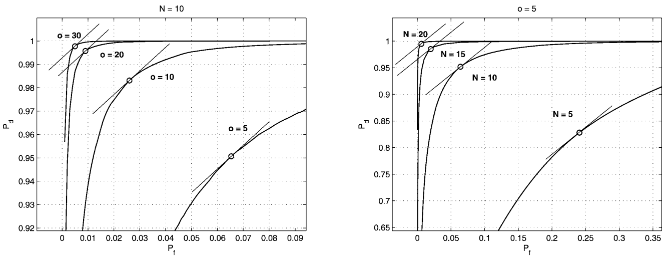
	o				
N	5	10	20	40	80
	$\lambda_N^{\uparrow}(\overline{\mathbf{y}}, o\mathbf{1})$				
3	5.27	5.93	6.00	6.00	6.00
5	33.88	71.59	104.38	117.30	119.87
	$\lambda_N^{\uparrow}(\underline{\mathbf{y}}, o\mathbf{1})$				
3	1.62e-004	1.55e-008	1.09e-016	3.93e-033	3.71e-066
5	1.16e-007	7.69e-015	2.58e-029	2.12e-058	1.04e-116
	$\Lambda_N(\breve{\mathbf{y}}, o\mathbf{1})$				
3	6.17e-003	8.47e-006	1.11e-011	1.41e-023	1.64e-047
5	4.39e-005	9.14e-011	1.71e-022	2.91e-046	4.08e-094

of possible DPMLRs increases. Furthermore, if $\mathbf{y}$ happens to exhibit a perfect monotonic ordering, then the evidence to support H_1 also increases as the spread of the y_i's increases. In other words, the chances of measurement errors leading to errors in inferring the wrong hypothesis decreases as the number of observers increases. The performance of the DPMLRT in terms of hypothesis errors is evaluated by Monte Carlo simulations in the next section.



(a) (b)

Fig. 2. ROC curves for DPMLR, MC, Spearman correlation, Kendall correlation and logistic correlation tests. (a) $o = 10$. (b) $o = 30$.



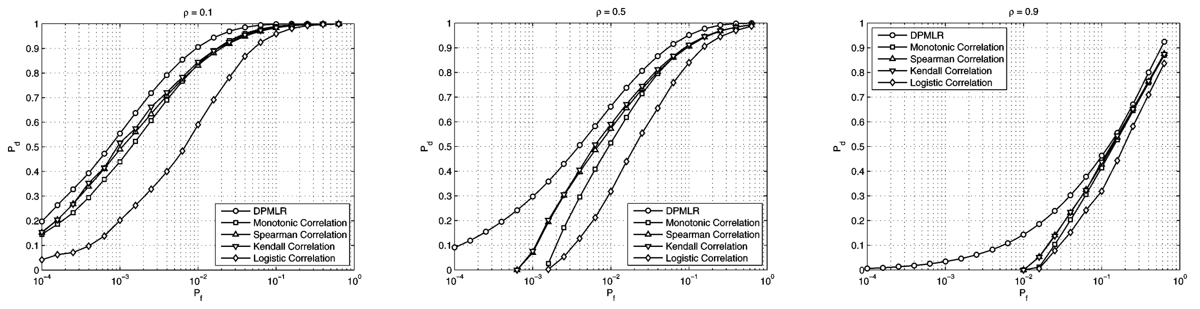
(a) (b)

Fig. 3. ROC curves for DPMLRT for various values of N and o. (a) $N = 10$, and $o = 5, 10, 20$, or 30. (b) $o = 5$ and $N = 5, 10, 15$, or 20.

III. DPMLRT PERFORMANCE ANALYSIS

In this section, we justify the performance of the proposed DPMLRT. To this end, we generate Monte Carlo realizations of $\mathbf{y}$, $\tilde{\mathbf{x}}$, and $\tilde{\mathbf{p}}$. Specifically, the $\tilde{p}_i$'s are generated uniformly over $[0, 1)$. For the monotonic hypothesis, $\tilde{x}_i = (\tilde{p}_i)^{\alpha}$. For the null hypothesis, the $\tilde{x}_i$'s are i.i.d. from a uniform distribution. For either hypothesis, the y_i's are random realizations of the binomial distribution (see (5)). For a given hypothesis and values of $o\mathbf{1}$, N, and α, we generated 10^6 realizations of $\mathbf{y}$, $\tilde{\mathbf{x}}$, and $\tilde{\mathbf{p}}$, and we computed the associated DPMLR given one scene, i.e., $S = 1$. Then, we use the histograms of the DPMLR to generate ROC curves by varying the acceptance threshold and tabulating the number of acceptances under the monotonic hypothesis, i.e., probability of detection (P_d), and under the null hypothesis, i.e., probability of false alarms (P_f). As a means of comparison, we also compute ROC curves associated with some other correlation tests in a similar fashion over the same simulations.

Fig. 1 includes ROC curves of the DPMLR, the MC [17], the Pearson correlation and the logistic correlation [23], [27] tests for various values of α when $N = 10$ and $o = 5$. Interested readers are referred to [17] for a detailed description of the monotonic and logistic correlations. In Fig. 1(a), where $\alpha = 1$, the Pearson correlation performs better than the others. This is explained by the fact that the relationship between $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{p}}$ is actually linear, and Pearson correlation exploits the actual values of $\tilde{\mathbf{x}}$ and not just the ordering. In essence, the test for linearity is better in this case than the more general test of monotonicity because it exploits more information. As the $g(x)$ function becomes more nonlinear (i.e., α increases), the performance of the Pearson correlation degrades significantly. Clearly, the logistic correlation is more robust to the nonlinearity than the Pearson



(a) (b) (c)

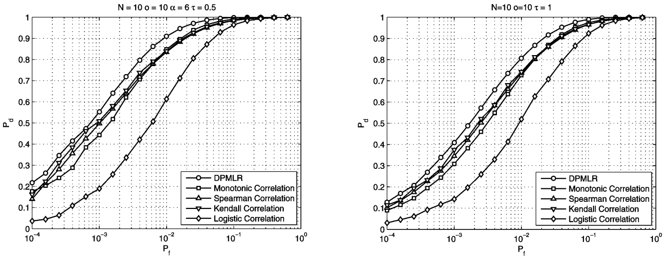
Fig. 4. ROC curves for correlated $\tilde{p}_i$'s. (a) $\rho = 0.1$. (b) $\rho = 0.5$. (c) $\rho = 0.9$.

correlation, but since not all monotonic relations follow a logistic function, the logistic correlation performs worse than the MC. Note that α does not change the ordering of $\tilde{x}_i$'s. Therefore, the performance of the DPMLRT and the MC is invariant to the nonlinearity. The DPMLRT always outperforms the MC because for the case of the uniform prior on $\tilde{\mathbf{p}}$, the DPMLRT is the most powerful test of monotonicity.

We also consider some common rank-order correlations—the Spearman correlation and the Kendall correlation when $N = 10$ and $\alpha = 6$. The ROC curves for different tests are shown in Fig. 2. Fig. 2(a) corresponds to a case where $o = 30$ and Fig. 2(b) corresponds to a case where $o = 10$. The DPMLRT always outperforms the others as expected given that the $\tilde{\mathbf{p}}$'s are generated by the assumed prior distribution. As the number of observers increases, the gap between the ROC curves of the DPMLRT and the rank-order correlation tests becomes larger. When $o = 10$, the performance of the MC is a little poorer than that of the rank-order correlations. For larger $o(o = 30)$, the MC outperforms the rank-order correlations because it takes advantage of the values of y_i's while the rank-order correlations only use their rank information. The logistic correlation exhibits the worst performance because of the limitation of the logistic regression fitting.

Fig. 3 provides the ROC curves of the DPMLRT for different o's and N's. The circle on each curve denotes the operating point when the threshold is set to one. As shown in [31], the slope of the ROC curve for a LRT is equal to the corresponding threshold value. Thus, when the threshold is one, the slope is one corresponding to the "knee" of the ROC curve as demonstrated in Fig. 3, which uses a linear scale for the P_f-axis. As one increases the number of observers, the knee of the ROC curve shifts to the top left corner, which means higher P_d and lower P_f can be achieved for a threshold of one. As expected, as the number of fused images N or the number of observers o increases, the efficacy of the DMPLR improves.

The next set of simulations consider how the DPMLR performs when the model assumptions do not match the data. For these simulations, $N = 10$, $o = 10$, and $\alpha = 6$. The first case considers uniform random variables $\tilde{p}_i$'s with a prespecified correlation matrix Σ, whose (m, n)th element denotes the correlation coefficient of $\tilde{p}_m$ and $\tilde{p}_n$ $(1 \leqslant m, n \leqslant N)$. The method for generating such $\tilde{p}_i$'s is from [32]. In this case we denote the nondiagonal elements of Σ by ρ (the diagonal elements equal 1). The $\tilde{p}_i$'s are completely correlated or independent for $\rho = 1$ or $\rho = 0$, respectively. Fig. 4 compares the ROC curves of DPMLR with the other correlations for different ρ's. Fig. 4(a)–(c) correspond to $\rho = 0.1, 0.5$ and 0.9, respectively. By comparing these ROC curves to Fig. 2, we can see that the gap between the DPMLRT and the others decreases as ρ increases. But clearly the DPMLRT exhibits the best performance among these correlations. In the limit, as ρ goes to 1, the monotonic evaluation is moot as all values of the $\tilde{p}_i$'s are equal.



(a) (b)

Fig. 5. ROC curves under generalized binomial distribution. (a) $\tau = 0.5$. (b) $\tau = 1$.

The next case considers the effect when the model of human performance does not match the binomial distribution. We consider the generalized binomial distribution [33] to incorporate diversity in the capabilities of humans. Specifically, the nominal human performance $\tilde{p}_i$ and associated FIQM $\tilde{x}_i = (\tilde{p}_i)^\alpha$ are generated as usual. Then, the realized mean performance for the observers $\hat{p}_i$ is drawn from the uniform distribution over $[\tilde{p}_i - \tau\tilde{p}_i(1 - \tilde{p}_i), \tilde{p}_i + \tau\tilde{p}_i(1 - \tilde{p}_i)]$ and y_i is drawn from a binomial distribution with parameters o and $\hat{p}_i$.[3] Here $\tau \in [0, 1]$ is referred to as the spread parameter, which denotes the deviation of y_i's distribution from the binomial distribution. Note that for $\tau = 0$, y_i still follows the binomial distribution with parameters o and $\tilde{p}_i$. Fig. 5 shows the ROC curves of the DPMLRT, the monotonic, the rank-order and the logistic correlation tests for different spread parameters τ. This figure demonstrates that the DPMLRT is robust to τ and still outperforms the others even when τ is as large as one.

The final case demonstrates that the DPMLR is not the UMP for any arbitrary prior distribution. Consider a pathological case

[3] As discussed in [33], any pmf of y_i over $[0, o]$ can be generated by choosing a specific pdf to generate $\hat{p}_i$.

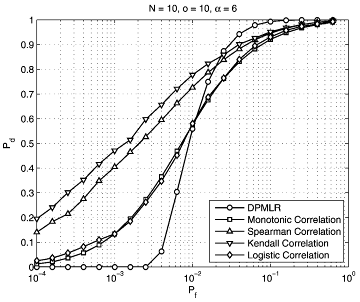


Fig. 6. Example of the case in which $\tilde{p}_i$'s are at the edge of $\mathcal{P}_\uparrow$ under H_1.

in which for odd i, $\tilde{p}_i = \tilde{p}_{i+1}$ and $\tilde{p}_i$ is drawn from a uniform distribution over $[0, 1)$. In practice, this case is unlikely because it means that two different fusion methods provide images with equivalent performance over multiple scenes. Nevertheless, Fig. 6 compares the DPMLRT with the other correlations. The figure shows that the DPMLRT outperforms the others when the P_d is high. But the other correlations all achieve higher detection probability than the DPMLRT when P_f is less than 0.01.

IV. FIQM Evaluation via the DPMLRT

This section demonstrates the application of the DPMLRT to score potential FIQMs. The DPMLRT and some other correlations are used to evaluate the monotonic relationships between 17 different FIQMs proposed in the literature and the human detection results in a specific target detection experiment. Details of the experiment and the discussion about the evaluation results are provided in the following subsections.

A. Experimental Setup

Long-wave infrared (LWIR) and image intensified (II) imagery were collected in a simulated military operation in an urban terrain (MOUT) environment. The imagery includes six interior and exterior locations, where four scenarios were collected for each location. The four scenarios represent cases where zero, one, two, and three people are within the field of regard of the camera. Individuals who were in the field of regard were typically obscured by objects in the scene, such as doorways, windows, furniture, and tables. For each of the scenarios, a horizontal pan of 150 images was then used to create a larger mosaic of imagery in both the LWIR and II bands.

The perceptual goal for the human observers is to detect the target in the scenes by interrogating the fused imagery. To generate the imagery, the LWIR and II images were registered, bore-sighted and fused via six different algorithms: 1) contrast pyramid A (CONA), 2) contrast pyramid B (CONB) [34], 3) discrete wavelet transform (DWTT) [1], [35], [36] 4) color discrete wavelet transform (CDWT), 5) color averaging (CLAV), and 6) color multiscale transform (CLMT) [37]. The first three algorithms generate grayscale fused images, and the final three methods generate color fused images. It is worth mentioning that the distinction between CONA and CONB is which image (LWIR or II) populates the coarsest coefficients in the pyramid. Also, the color methods generate a grayscale fusion method for the luminance component, map the differences in the image coefficients in the saturation component, and encode the source of the largest coefficient (LWIR versus II) in the hue component. The CDWT uses this coloring scheme for the DWT coefficients, the CLAV uses simple averaging for the luminance and the raw pixels for the color components, and the CLMT uses the coloring scheme for the multiscale fusion method defined in [37]. Finally, it is instructive to compare the fused imagery against the source imagery. Therefore, we consider eight fused image displays: 1) II, 2) LWIR, 3) CONA, 4) CONB, 5) DWTT, 6) CDWT, 7) CLAV, and 8) CLMT.

Fig. 7 shows an example of the resulting eight fused image displays for a typical scenario in our experiment.[4] In this scenario, there are two target persons which are highlighted by the boxes in each image. As seen in Fig. 7(b), the human targets stand out in the LWIR imagery because they are usually hotter than the background. For the most part, detection performance is best on the LWIR only band because the search task can often be reduced to simply finding the white hot object on a grey background. However, the II band has the potential to add context to the LWIR band as the objects like tables and chairs are easier to distinguish in the II band [see Fig. 7(a) and (b)]. Therefore, there can be value in fusing the two bands.

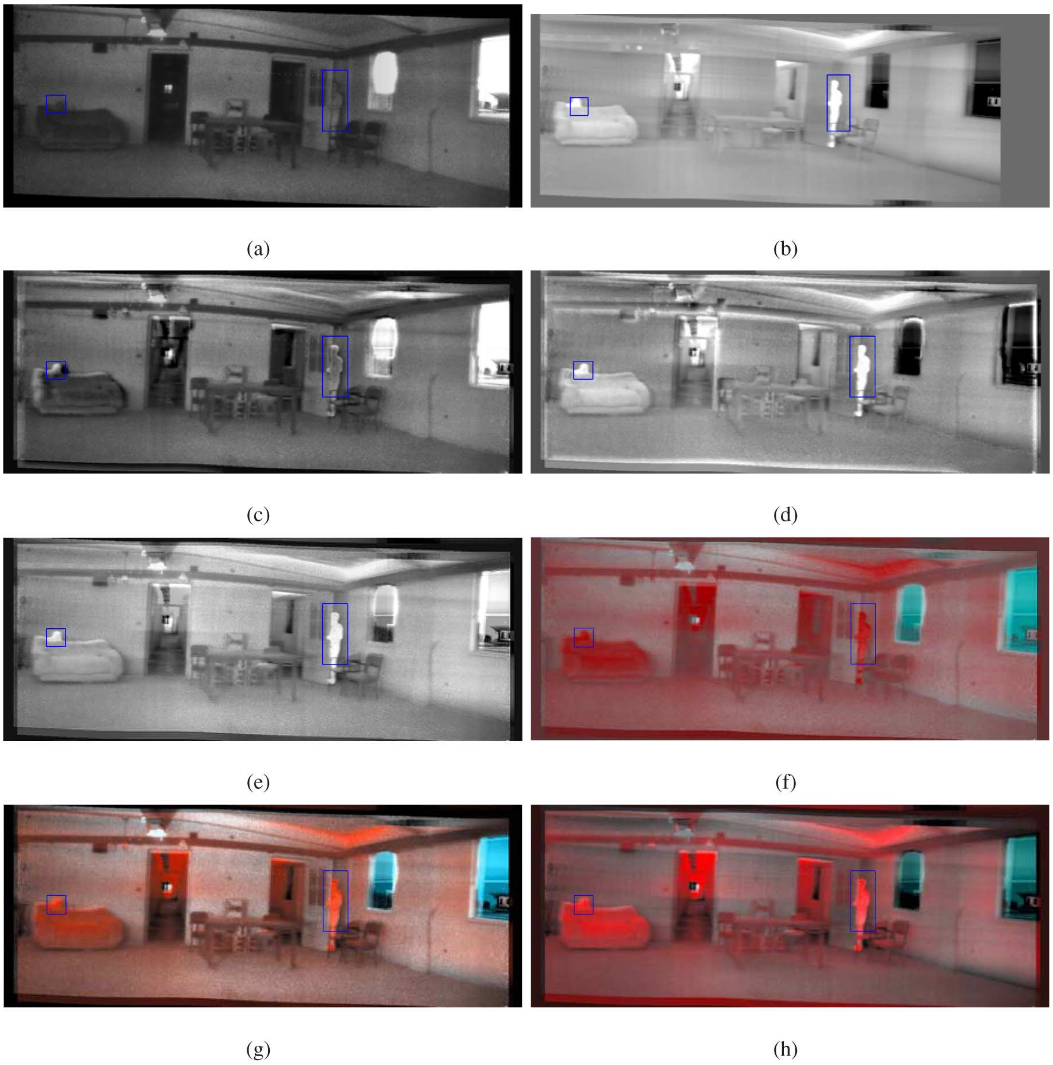
A perception test was set up whereby observers were asked to try to find the human targets in a "field of regard" search. An observer's display was calibrated to look as though it were seeing a single field of regard of a given scene, and the observer had to navigate across the scene and detect human targets. Observers could mark as many as three places on the display as detections for human targets (as they were told that the images could contain between zero and three humans hiding in the scene). At any point an observer could push a button to indicate that they either did not detect any targets in the scene or that there were no other targets in the scene. In the end, the detection performance of the humans was recorded over the eight image displays.

Overall, $o = 8$ observers evaluated 18 scenarios that contained 35 human targets. We treat each target and its surrounding area as a scene for every scenario. For example, the inside of each box in Fig. 7 represents a scene, as shown in Fig. 8(a)–(h). Then, $\mathbf{y}_s$ is the number of observers that correctly detected the target located in the sth scene for $s = 1, \ldots, 35$.

B. Evaluated FIQMs

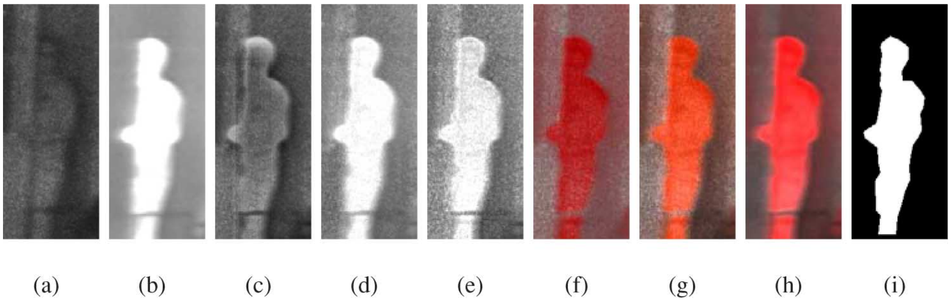
We test 17 potential FIQMs over each scene. These FIQMs are listed in Table II with corresponding citations. Most measures listed in Table II were also evaluated in [17] for a recognition task. All the measures except the first are computed automatically. The first ten measures are simply complexity features that do not consider the source images (the no-source comparative class according to the classification in Section I). They represent the structure, texture, contrast and/or edge intensities in the image in order to characterize the complexity of the image. Such measures have already been used to evaluate the quality of image fusion algorithms [17], [18], [38]. Most of these measures have been inspired by work to develop clutter complexity measures [19], [39]. These works search for features that characterize the degree to which the background appears target-like [39]. Ideally, the clutter complexity determines how hard it is to detect or classify a target in the scene due to the complexity of the background. The last seven measures compare how well the

[4] The color versions of the CDWT, CLAV, and CLMT displays in Figs. 7–8 are available in the online version of this paper.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 7. Eight fused image displays for one of the 18 scenarios: (a) II, (b) LWIR, (c) CONA, (d) CONB, (e) DWTT, (f) CDWT, (g) CLAV, and (h) CLMT.



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 8. Example of the eight fused image displays and corresponding silhouette for a single scene, i.e., a target instance, in Fig. 7: (a) II, (b) LWIR, (c) CONA, (d) CONB, (e) DWTT, (f) CDWT, (g) CLAV, (h) CLMT, and (i) silhouette.

salient features in the two source imagery are transferred into the fused image (the source comparative class). For the most part, the distinction between these comparative measures is in the definition of saliency.

Ideally, the FIQM should be computed automatically from the fused and source images. The contrast measure is considered because it is one of the measures that is averaged in an objective National Imagery Interpretability Ratings Scale (NIIRS) rating [41]. Furthermore, it is intuitive that the contrast between the target and the background facilitates ease of detection. The contrast is computed by manually segmenting human silhouettes for each scene. Fig. 8(i) shows an example of the silhouette that separates the target from the background. The white part in the silhouette denotes target pixels, and the black part denotes back-

TABLE II
LIST OF THE EVALUATED FIQMs

Category	Index	Measure Description
Contrast	1	Difference of intensity or color between the target and the background
Saturation [17]	2	Normalized histogram peak
STD	3	Standard deviation
Schmieder Weathersby [19]	4	Block average local standard deviation
fBm [20]	5	Hurst parameter for fBm model
TIR [21]	6	Block average target interference ratio (contrast)
Energy [21]	7	Block average energy of histogram
Entropy [21]	8	Block average entropy of histogram
Homogeneity [21]	9	Block average pixel variation
Block Outlier [21]	10	Block average number of outliers
Universal Quality Index [23]	11	Average Structure SIMilarity (SSIM) index between fused and reference images
Information Measures [11]	12	Average mutual information between fused and reference images (bin size = 16)
Objective Measure [10]	13	Average objective edge information between fused and reference images
Salient Quality Index [12]	14	Weighted average salient quality index of edge intensities between fused and reference images
	15	Weighted average salient quality index between fused and reference images
	16	Average salient quality index between fused and reference images
Harris Response based quality metric [40]	17	Difference of Harris response between fused and reference images

ground pixels. The measure is equivalent to the percent contrast used in [42]. For grayscale imagery, it is defined as

$$\text{contrast} = \frac{|I_t - I_b|}{d} \tag{23}$$

where I_t and I_b are the mean target and background intensities, respectively, and d denotes the dynamic range, i.e., the intensity difference between the brightest and darkest pixels in a scene. For color imagery, the RGB coordinates are converted to the CIE $L^*a^*b^*$ color space [43] and the single band contrast is calculated independently over the L^*, a^*, and b^* bands via (23). Then the root sum square of the three single band contrasts is reported as the overall contrast. Since the information about the color is given in the a^* and b^* bands, these bands exhibit zero contrast for grayscale imagery, and the color version of contrast is a consistent generalization of the grayscale definition, i.e., it provides the same answer if the RGB image contains no color. Intuitively, the color version of contrast integrates the contrast that exists in all ways the eye can distinguish the foreground from the background, i.e., lightness and color. It might be possible to generate an automated contrast measure by incorporating automated image segmentation techniques. This is a matter of future investigation.

While the generalization of contrast for color imagery is straightforward, it is not clear how to best extend the definition of the other automatic FIQMs to accommodate color imagery. To this end, we follow the convention in [39] where for the color images, one generates four color measures for a given grayscale measure. Namely, the grayscale measure is computed over each RGB band and summarized by the 1) maximum, 2) minimum, and 3) median values over all bands. The fourth measure is computed by converting the RGB image into a grayscale image before calculating the measure.

C. Evaluation Results and Discussion

First, we evaluated the consistency of the FIQMs with human detection performance over the five grayscale fused image displays: 1) II, 2) LWIR, 3) CONA, 4) CONB, and 5) DWTT. Then, we considered scoring the FIQMs generalized for color using all eight fused image displays.

TABLE III
LIST OF DPMLR SCORES, ASSOCIATED p-VALUES, AND AVERAGE VALUES OF SOME OTHER CORRELATIONS FOR 17 GRAYSCALE FIQMS TESTED OVER FIVE IMAGE DISPLAYS

Index	Λ_N	p-value	Mean MC	Mean $\|MC\|$	Mean LC	Mean $\|LC\|$	Mean SC	Mean $\|SC\|$	Mean KC	Mean $\|KC\|$
1	1.3291	0.1221	0.5835	0.8276	0.5628	0.8069	0.5210	0.6168	0.4700	0.5602
2	0.0204	0.4284	0.0153	0.6962	0.0105	0.6382	0.0053	0.4158	0.0190	0.3425
3	0.0301	0.3982	0.1291	0.7958	0.1374	0.7714	0.0868	0.4981	0.0891	0.4285
4	0.0340	0.3884	0.0192	0.7965	0.0180	0.7911	0.0720	0.4498	0.0711	0.3807
5	0.5925	0.1741	0.5322	0.8564	0.5330	0.8398	0.3608	0.5730	0.3094	0.4933
6	0.3637	0.2073	0.1380	0.5493	0.1381	0.5490	0.1381	0.5490	0.0649	0.4158
7	0.0376	0.3803	0.0546	0.7670	0.0405	0.7248	0.0748	0.4433	0.0729	0.3780
8	0.0392	0.3768	0.0921	0.7636	0.0883	0.7386	0.0896	0.4434	0.0745	0.3730
9	0.0382	0.3789	-0.0106	0.8068	-0.0012	0.7851	-0.0360	0.5160	-0.0292	0.4384
10	0.0422	0.3709	0.1453	0.7723	0.1385	0.7555	0.0927	0.5002	0.0807	0.4197
11	0.0316	0.3943	-0.0292	0.7543	-0.0268	0.7339	0.0130	0.4066	0.0345	0.3494
12	0.0362	0.3833	0.2030	0.7262	0.1667	0.6771	0.1363	0.3950	0.1315	0.3457
13	0.0479	0.3607	0.1018	0.7863	0.1063	0.7622	0.0647	0.4677	0.0651	0.4094
14	0.0252	0.4120	0.1454	0.7668	0.1426	0.7581	0.0792	0.4340	0.0870	0.3743
15	0.0242	0.4150	0.1541	0.7756	0.1402	0.7558	0.0873	0.4421	0.0934	0.3807
16	0.0387	0.3778	0.3124	0.7694	0.2969	0.7391	0.1534	0.4396	0.1430	0.3898
17	0.3407	0.2122	0.1973	0.5980	0.1931	0.5910	0.1424	0.3792	0.1427	0.3527

Table III provides the composite DPMLR score over the five grayscale displays of the 35 scenes for each of the 17 grayscale measures as well as the corresponding p-values. Note that for each FIQM, the p-value is evaluated by calculating the probability of obtaining a result with the DPMLR larger than the composite DPMLR score listed in Table III when the H_0 hypotheses is true. Furthermore, the table also includes the average values and average absolute values of the monotonic, logistic, Spearman, and Kendall correlations.

The second column of Table III shows that the composite DPMLR scores for all but the grayscale contrast measure are significantly less than one. This means that the evidence points to the fact that these potential FIQMs are viewed as noise with respect to ordering the detection probabilities of the imagery. The poor performance of the source comparative measures may be explained by structure in the fused and source images that leads to good interimage correlation but that has no (or even negative) effect on human performance. Examples of the pitfalls of source comparative measures when the ideal image is unknown are provided in [14].

For the grayscale contrast measure, the composite DPMLR score is still modest at 1.3291 and the p-value is not very low. In fact, the perfect FIQM that consistently ordered the number of detections $\mathbf{y}$ over all 35 scenes would provide a composite DPMLR of 9.632. This means that while there is evidence to reject the null hypothesis, the evidence to support the monotonic hypothesis is not compelling. However, the composite DPMLR score for the grayscale contrast measure is much greater than the scores for the others. Thus, the contrast measure may be a key aspect to a proper FIQM.

From Table III, one can see that the orderings of the FIQMs via the DPMLR and the other correlations differ. Also note that for each FIQM, the differences between the average correlations and the average absolute correlations indicate a consistency issue for the nature of monotonicity over the 35 scenes. The contrast measure exhibits by far the largest DPMLR. However, its average absolute values of the MC and the logistic correlation (mean $|\mathrm{MC}|$ and mean $|\mathrm{LC}|$ in Table III) are less than those of the fBm, respectively. Furthermore, the other average correlations of the contrast measure are only slightly larger than those of the fBm. To better compare these two measures, and to show how differently the DPMLR and the other four correlations evaluate a FIQM based upon the human perception results, we present the human detection results and the scores of the DPMLRT, monotonic, logistic, Spearman, and Kendall correlation tests for each scene for the contrast and the fBm measures.

Fig. 9 graphically depicts the relationship between the aforementioned two measures and the human performance over all 35 scenes. The lines marked by the asterisk correspond to the contrast measure and the lines marked by the circle correspond to the fBm. Since only the five gray fused image displays are considered here, for each scene and each FIQM, we have five detection numbers $y_i \in [0, 8]$ and five FIQM values $\tilde{x}_i$ $(1 \leqslant i \leqslant 5)$. In each plot of Fig. 9, the vertical axis denotes the number of humans that detected the target, while the horizontal axis stands for the rank of the $\tilde{x}_i$'s sorted in ascending order. The shade of the background of each plot indicates the significance of the monotonic ordering for each scene. The significance value is obtained by calculating the DPMLR of the given y_i's for an imaginary FIQM whose values perfectly match the y_i's in the monotonically increasing order.

Tables IV and V provide the ascending and descending DPMLRs as well as the other four correlations (monotonic, logistic, Spearman, and Kendall) over each scene for the contrast measure and the fBm measure, respectively. Note that in Scenes 34 and 35, the same number of detections are obtained for five different displays. Because of the fact that the target is so obvious in Scene 34, all eight observers detected it successfully. Similarly, no one detected the target in Scene 35 because it is so unclear. Both cases are naturally ignored as they don't provide any information on the monotonicity.

One very important property of the DPMLR is that it can capture the significance of a scene based upon the human detection results, and accordingly adjust its score to provide a more precise evaluation. The significance, as defined, is determined by the number of unique human detection values and the spread of these values over the dynamic range from 0 to 8 detections. Essentially, the significance describes how easy (or difficult) it

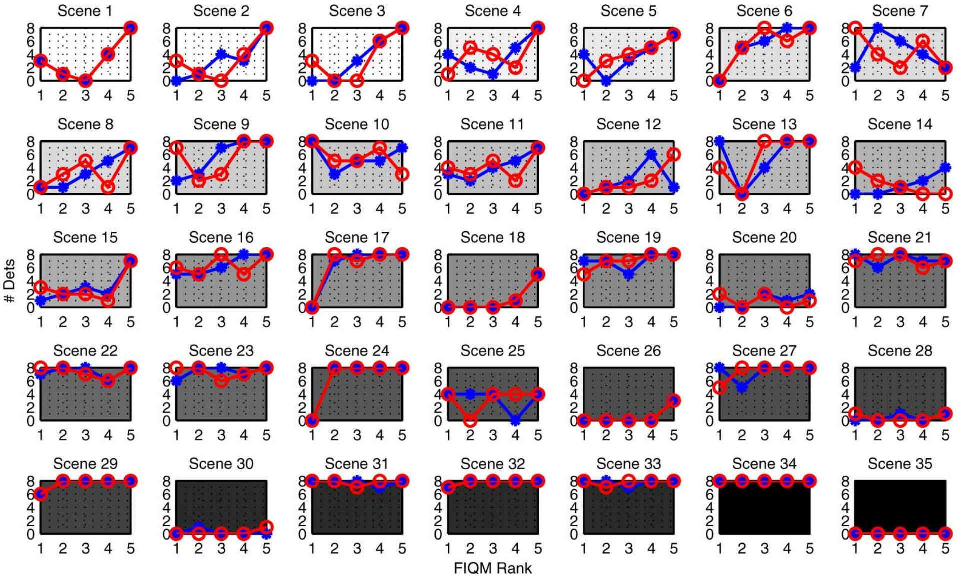


Fig. 9. Scatter plots of the number of detections versus the quality rank order over 35 scenes ("$*$" represents the contrast measure, and "$\circ$" represents the fBm measure).

TABLE IV
STATISTICS FOR THE CONTRAST MEASURE WHERE MC, LC, SC, AND KC ARE THE MONOTONIC, LOGISTIC, SPEARMAN, AND KENDALL CORRELATIONS, RESPECTIVELY

Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Scene 6	Scene 7
$\lambda_5^{\uparrow} = 0.7043$	$\lambda_5^{\uparrow} = 23.6503$	$\lambda_5^{\uparrow} = 45.2298$	$\lambda_5^{\uparrow} = 1.1634$	$\lambda_5^{\uparrow} = 0.6730$	$\lambda_5^{\uparrow} = 31.4151$	$\lambda_5^{\uparrow} = 0.0010$
$\lambda_5^{\downarrow} = 0.0000$	$\lambda_5^{\downarrow} = 0.0000$	$\lambda_5^{\downarrow} = 0.0000$	$\lambda_5^{\downarrow} = 0.0000$	$\lambda_5^{\downarrow} = 0.0001$	$\lambda_5^{\downarrow} = 0.0000$	$\lambda_5^{\downarrow} = 0.0137$
MC = 0.9379	MC = 0.9935	MC = 1.0000	MC = 0.9189	MC = 0.8376	MC = 1.0000	MC = -0.5601
LC = 0.9379	LC = 0.9699	LC = 0.9990	LC = 0.9189	LC = 0.8376	LC = 1.0000	LC = -0.5601
SC = 0.6000	SC = 0.9000	SC = 0.9747	SC = 0.6000	SC = 0.7000	SC = 0.9747	SC = -0.2052
KC = 0.4000	KC = 0.8000	KC = 0.9487	KC = 0.4000	KC = 0.6000	KC = 0.9487	KC = -0.3162
Scene 8	Scene 9	Scene 10	Scene 11	Scene 12	Scene 13	Scene 14
$\lambda_5^{\uparrow} = 29.1496$	$\lambda_5^{\uparrow} = 24.7705$	$\lambda_5^{\uparrow} = 0.0274$	$\lambda_5^{\uparrow} = 11.1297$	$\lambda_5^{\uparrow} = 0.2919$	$\lambda_5^{\uparrow} = 0.0005$	$\lambda_5^{\uparrow} = 17.3610$
$\lambda_5^{\downarrow} = 0.0000$	$\lambda_5^{\downarrow} = 0.0000$	$\lambda_5^{\downarrow} = 0.0746$	$\lambda_5^{\downarrow} = 0.0014$	$\lambda_5^{\downarrow} = 0.0006$	$\lambda_5^{\downarrow} = 0.0000$	$\lambda_5^{\downarrow} = 0.0009$
MC = 1.0000	MC = 1.0000	MC = -0.6882	MC = 0.9830	MC = 0.6571	MC = 0.6124	MC = 1.0000
LC = 0.9826	LC = 0.8778	LC = -0.6882	LC = 0.9558	LC = 0.5026	LC = 0.6124	LC = 1.0000
SC = 0.9747	SC = 0.9747	SC = -0.0513	SC = 0.9000	SC = 0.5643	SC = 0.3354	SC = 0.9747
KC = 0.9487	KC = 0.9487	KC = 0.1054	KC = 0.8000	KC = 0.5270	KC = 0.3586	KC = 0.9487
Scene 15	Scene 16	Scene 17	Scene 18	Scene 19	Scene 20	Scene 21
$\lambda_5^{\uparrow} = 9.4196$	$\lambda_5^{\uparrow} = 11.8403$	$\lambda_5^{\uparrow} = 10.8893$	$\lambda_5^{\uparrow} = 10.4186$	$\lambda_5^{\uparrow} = 1.0475$	$\lambda_5^{\uparrow} = 4.7336$	$\lambda_5^{\uparrow} = 0.1784$
$\lambda_5^{\downarrow} = 0.0002$	$\lambda_5^{\downarrow} = 0.0022$	$\lambda_5^{\downarrow} = 0.0000$	$\lambda_5^{\downarrow} = 0.0001$	$\lambda_5^{\downarrow} = 0.0376$	$\lambda_5^{\downarrow} = 0.0250$	$\lambda_5^{\downarrow} = 0.7175$
MC = 0.9886	MC = 1.0000	MC = 1.0000	MC = 1.0000	MC = 0.7454	MC = 0.9354	MC = -0.5345
LC = 0.9708	LC = 1.0000	LC = 1.0000	LC = 1.0000	LC = 0.7454	LC = 0.8815	LC = -0.5345
SC = 0.8208	SC = 0.9487	SC = 0.8944	SC = 0.8944	SC = 0.6325	SC = 0.7906	SC = -0.2635
KC = 0.7379	KC = 0.8944	KC = 0.8367	KC = 0.8367	KC = 0.4472	KC = 0.6708	KC = -0.2236
Scene 22	Scene 23	Scene 24	Scene 25	Scene 26	Scene 27	Scene 28
$\lambda_5^{\uparrow} = 0.3496$	$\lambda_5^{\uparrow} = 1.7957$	$\lambda_5^{\uparrow} = 4.9996$	$\lambda_5^{\uparrow} = 0.0108$	$\lambda_5^{\uparrow} = 4.3533$	$\lambda_5^{\uparrow} = 0.5042$	$\lambda_5^{\uparrow} = 2.0380$
$\lambda_5^{\downarrow} = 0.2204$	$\lambda_5^{\downarrow} = 0.0501$	$\lambda_5^{\downarrow} = 0.0000$	$\lambda_5^{\downarrow} = 0.1740$	$\lambda_5^{\downarrow} = 0.0042$	$\lambda_5^{\downarrow} = 0.0259$	$\lambda_5^{\downarrow} = 0.1714$
MC = -0.4082	MC = 0.8898	MC = 1.0000	MC = -0.6124	MC = 1.0000	MC = 0.6124	MC = 0.7638
LC = -0.4082	LC = 0.8458	LC = 1.0000	LC = -0.6124	LC = 1.0000	LC = 0.6124	LC = 0.5417
SC = 0.1118	SC = 0.4472	SC = 0.7071	SC = -0.3536	SC = 0.7071	SC = 0.3536	SC = 0.5774
KC = 0.1195	KC = 0.3586	KC = 0.6325	KC = -0.3162	KC = 0.6325	KC = 0.3162	KC = 0.5164
Scene 29	Scene 30	Scene 31	Scene 32	Scene 33	Scene 34	Scene 35
$\lambda_5^{\uparrow} = 3.5970$	$\lambda_5^{\uparrow} = 0.4156$	$\lambda_5^{\uparrow} = 0.4156$	$\lambda_5^{\uparrow} = 2.4100$	$\lambda_5^{\uparrow} = 0.7393$	$\lambda_5^{\uparrow} = 1.0000$	$\lambda_5^{\uparrow} = 1.0000$
$\lambda_5^{\downarrow} = 0.0296$	$\lambda_5^{\downarrow} = 1.2533$	$\lambda_5^{\downarrow} = 1.2533$	$\lambda_5^{\downarrow} = 0.1818$	$\lambda_5^{\downarrow} = 0.7393$	$\lambda_5^{\downarrow} = 1.0000$	$\lambda_5^{\downarrow} = 1.0000$
MC = 1.0000	MC = -0.6124	MC = -0.6124	MC = 1.0000	MC = 0.4082	MC = NaN	MC = NaN
LC = 1.0000	LC = -0.6124	LC = -0.6124	LC = 1.0000	LC = 0.4082	LC = NaN	LC = NaN
SC = 0.7071	SC = -0.3536	SC = -0.3536	SC = 0.7071	SC = 0.0000	SC = NaN	SC = NaN
KC = 0.6325	KC = -0.3162	KC = -0.3162	KC = 0.6325	KC = 0.0000	KC = NaN	KC = NaN

is for random noise to affect the order of the human detection results. The more unique values that the human detection results take in a scene, the less likely that random noise will order the human detection results. The monotonic and logistic correlations give a value of one whenever a scene's scatter plot is perfectly monotonic, as observed from Scenes 3, 6, 8, 9, 14, 16, 17, 18, 24, 26, 29, and 32 in Fig. 9 and the corresponding statistics in Table IV. The Spearman and the Kendall correlations give a value of one whenever a scene's scatter plot is strictly monotonic, as observed from the scatter plot of Scene 5 and the

TABLE V
STATISTICS FOR THE FBM WHERE MC, LC, SC AND KC ARE THE MONOTONIC, LOGISTIC, SPEARMAN, AND KENDALL CORRELATIONS, RESPECTIVELY

Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Scene 6	Scene 7
$\lambda_5^{\uparrow}$ = 0.7043	$\lambda_5^{\uparrow}$ = 0.7043	$\lambda_5^{\uparrow}$ = 0.3819	$\lambda_5^{\uparrow}$ = 1.5343	$\lambda_5^{\uparrow}$ = 35.3484	$\lambda_5^{\uparrow}$ = 6.8845	$\lambda_5^{\uparrow}$ = 0.0001
$\lambda_5^{\downarrow}$ = 0.0000	$\lambda_5^{\downarrow}$ = 0.0000	$\lambda_5^{\downarrow}$ = 0.0000	$\lambda_5^{\downarrow}$ = 0.0000	$\lambda_5^{\downarrow}$ = 0.0000	$\lambda_5^{\downarrow}$ = 0.0000	$\lambda_5^{\downarrow}$ = 1.0015
MC = 0.9379	MC = 0.9379	MC = 0.9396	MC = 0.9189	MC = 1.0000	MC = 0.9766	MC = -0.8402
LC = 0.9379	LC = 0.9379	LC = 0.9396	LC = 0.8385	LC = 0.9766	LC = 0.9686	LC = -0.7796
SC = 0.6000	SC = 0.6000	SC = 0.6669	SC = 0.6000	SC = 1.0000	SC = 0.8208	SC = -0.6156
KC = 0.4000	KC = 0.4000	KC = 0.5270	KC = 0.4000	KC = 1.0000	KC = 0.7379	KC = -0.5270
Scene 8	Scene 9	Scene 10	Scene 11	Scene 12	Scene 13	Scene 14
$\lambda_5^{\uparrow}$ = 0.8749	$\lambda_5^{\uparrow}$ = 0.0708	$\lambda_5^{\uparrow}$ = 0.0011	$\lambda_5^{\uparrow}$ = 0.6460	$\lambda_5^{\uparrow}$ = 19.4999	$\lambda_5^{\uparrow}$ = 0.2934	$\lambda_5^{\uparrow}$ = 0.0009
$\lambda_5^{\downarrow}$ = 0.0001	$\lambda_5^{\downarrow}$ = 0.0000	$\lambda_5^{\downarrow}$ = 2.5692	$\lambda_5^{\downarrow}$ = 0.0064	$\lambda_5^{\downarrow}$ = 0.0001	$\lambda_5^{\downarrow}$ = 0.0000	$\lambda_5^{\downarrow}$ = 17.3610
MC = 0.8402	MC = 0.7605	MC = -0.9081	MC = 0.8137	MC = 1.0000	MC = 0.9186	MC = -1.0000
LC = 0.7956	LC = 0.7605	LC = -0.8147	LC = 0.8137	LC = 0.9939	LC = 0.9186	LC = -0.9801
SC = 0.5643	SC = 0.6669	SC = -0.6669	SC = 0.3000	SC = 0.9747	SC = 0.7826	SC = -0.9747
KC = 0.5270	KC = 0.5270	KC = -0.5270	KC = 0.2000	KC = 0.9487	KC = 0.5976	KC = -0.9487
Scene 15	Scene 16	Scene 17	Scene 18	Scene 19	Scene 20	Scene 21
$\lambda_5^{\uparrow}$ = 0.7544	$\lambda_5^{\uparrow}$ = 0.5725	$\lambda_5^{\uparrow}$ = 5.2378	$\lambda_5^{\uparrow}$ = 10.4186	$\lambda_5^{\uparrow}$ = 9.9553	$\lambda_5^{\uparrow}$ = 0.0779	$\lambda_5^{\uparrow}$ = 0.1318
$\lambda_5^{\downarrow}$ = 0.0007	$\lambda_5^{\downarrow}$ = 0.0099	$\lambda_5^{\downarrow}$ = 0.0000	$\lambda_5^{\downarrow}$ = 0.0001	$\lambda_5^{\downarrow}$ = 0.0071	$\lambda_5^{\downarrow}$ = 0.5866	$\lambda_5^{\downarrow}$ = 0.8082
MC = 0.9535	MC = 0.6757	MC = 0.9949	MC = 1.0000	MC = 1.0000	MC = -0.6124	MC = -0.7638
LC = 0.9535	LC = 0.6233	LC = 0.9897	LC = 1.0000	LC = 0.9576	LC = -0.5001	LC = -0.7638
SC = 0.0513	SC = 0.3162	SC = 0.6708	SC = 0.8944	SC = 0.9487	SC = -0.3162	SC = -0.3689
KC = -0.1054	KC = 0.2236	KC = 0.5976	KC = 0.8367	KC = 0.8944	KC = -0.2236	KC = -0.2236
Scene 22	Scene 23	Scene 24	Scene 25	Scene 26	Scene 27	Scene 28
$\lambda_5^{\uparrow}$ = 0.1434	$\lambda_5^{\uparrow}$ = 0.2252	$\lambda_5^{\uparrow}$ = 4.9996	$\lambda_5^{\uparrow}$ = 0.1740	$\lambda_5^{\uparrow}$ = 4.3533	$\lambda_5^{\uparrow}$ = 4.3533	$\lambda_5^{\uparrow}$ = 0.4822
$\lambda_5^{\downarrow}$ = 1.0109	$\lambda_5^{\downarrow}$ = 0.6329	$\lambda_5^{\downarrow}$ = 0.0000	$\lambda_5^{\downarrow}$ = 0.0108	$\lambda_5^{\downarrow}$ = 0.0042	$\lambda_5^{\downarrow}$ = 0.0042	$\lambda_5^{\downarrow}$ = 0.4822
MC = -0.6124	MC = -0.6124	MC = 1.0000	MC = 0.6124	MC = 1.0000	MC = 1.0000	MC = 0.6124
LC = -0.6124	LC = -0.6124	LC = 1.0000	LC = 0.6124	LC = 1.0000	LC = 1.0000	LC = 0.6124
SC = -0.3354	SC = -0.2236	SC = 0.7071	SC = 0.3536	SC = 0.7071	SC = 0.7071	SC = 0.0000
KC = -0.3586	KC = -0.1195	KC = 0.6325	KC = 0.3162	KC = 0.6325	KC = 0.6325	KC = 0.0000
Scene 29	Scene 30	Scene 31	Scene 32	Scene 33	Scene 34	Scene 35
$\lambda_5^{\uparrow}$ = 3.5970	$\lambda_5^{\uparrow}$ = 2.4100	$\lambda_5^{\uparrow}$ = 0.7393	$\lambda_5^{\uparrow}$ = 2.4100	$\lambda_5^{\uparrow}$ = 1.2533	$\lambda_5^{\uparrow}$ = 1.0000	$\lambda_5^{\uparrow}$ = 1.0000
$\lambda_5^{\downarrow}$ = 0.0296	$\lambda_5^{\downarrow}$ = 0.1818	$\lambda_5^{\downarrow}$ = 0.7393	$\lambda_5^{\downarrow}$ = 0.1818	$\lambda_5^{\downarrow}$ = 0.4156	$\lambda_5^{\downarrow}$ = 1.0000	$\lambda_5^{\downarrow}$ = 1.0000
MC = 1.0000	MC = 1.0000	MC = 0.4082	MC = 1.0000	MC = 0.6124	MC = NaN	MC = NaN
LC = 1.0000	LC = 1.0000	LC = 0.4082	LC = 1.0000	LC = 0.6124	LC = NaN	LC = NaN
SC = 0.7071	SC = 0.7071	SC = 0.0000	SC = 0.7071	SC = 0.3536	SC = NaN	SC = NaN
KC = 0.6325	KC = 0.6325	KC = 0.0000	KC = 0.6325	KC = 0.3162	KC = NaN	KC = NaN

corresponding statistics in Table V. However, the DPMLR gives a score much greater than one in the significant scenes. Specifically, in Scenes 3, 6, 8, 9, and 14, the likelihoods for noise to order the data are much slimmer than those in Scenes 24, 26, 29, and 32. As a result, the DPMLR provides significantly higher scores in the former than in the later as seen in Table IV.

One can also observe that the miss-ordering for the more significant scenes causes lower DPMLR scores than those of the less significant scenes. For instance, we compare the scatter plots of Scenes 25 and 31 in Fig. 9 as well as the corresponding descending DPMLR values in Table IV. The descending DPMLR for Scene 25 is much smaller than that of Scene 31 because the DPMLR treats the miss-ordering in Scene 31 as due to the measurement noise. On the other hand, the other four correlations are equally unforgiving of the miss-ordering regardless of the significance of the scene. This is because that the correlations are invariant to linear scaling of the human detection results, whereas the DPMLR uses the binomial measurement model to determine whether or not the scale of the miss-ordering is significant.

Once we realize the DPMLR's ability to incorporate the significance of each scene into the statistical test, it is easy to see why the DPMLR provides the significantly high score for the contrast measure. Comparing the scatter plots of these two measures in the first 20 scenes, we can see that seven of them, i.e., Scenes 3, 6, 8, 9, 14, 16, and 17, exhibit the perfect monotonicity for the contrast measure and some miss-orderings for the fBm measure. In fact, the nature of the monotonic relationship for the fBm feature flips for Scene 14, i.e., it is perfectly decreasing. Both of these factors lead to the significantly higher DPMLR for the contrast measure. Because the contrast measure is still not nearly monotonically related to the perception results of many of the significant scenes, the composite DPMLR score is only slightly greater than one.

Next, we used all $N = 8$ fused image displays and ran another DPMLRT for color-based FIQMs when the human detection results were collected over $o = 8$ observers. The composite DPMLR scores of the 64 color-based FIQMs derived from the 16 automated grayscale FIQMs are low and not included here for the sake of brevity. On the other hand, the color-based contrast measure achieved a composite DPMLR score of 1.4000, which is slightly greater than that of the contrast computed only over the five grayscale fused image displays. Because the number of fused images N has increased, the significance of this "greater than one" score increases and the p-value is 0.041972, which gives stronger support for the monotonic hypothesis. Certainly, the color-based contrast is able to incorporate the contrast from both the luminance and color components in an RGB image and serves as a potential FIQM that is able to explain some of the human performance. Again, the perfect FIQM would provide a composite DPMLR of 54.5150, and contrast is only one aspect of a good FIQM, which has yet to be identified.

V. Conclusion

This paper proposes the composite DPMLR to quantify how consistent the values of a FIQM are with measured human performance represented by the probability of detection. Specifically, the DPMLR can be used to test whether or not a monotonic relationship exists between the FIQM and the underlying human detection performance that is measured via a perception experiment. The resulting test is designed to be applicable even when the number of observers is small so that the measurement errors from the perceptual experiment are not necessarily Gaussian. The paper discusses some interesting properties of the DPMLR, and simulation results demonstrate the advantages of the DPMLR over other monotonic statistics. Unlike the MC in [17], the DPMLR seamlessly accounts for the spread of the human observations and the number of fused images. It indicates to what degree the ordering of the human observations by the FIQM is not by random chance. The DPMLRT is a general test of monotonicity that can be used to evaluate monotonic relationships beyond the image fusion application. Finally, the DPMLR was used to score a number of potential FIQMs using real image data with a corresponding perception study.

The DPMLR scores reveal that a proper FIQM for the detection task is not yet available. The comparative measures may have scored poorly because the salient features exploited by these measures may not have captured the context in II imagery that humans exploit for detection. Of note, the contrast measure does demonstrate some utility based upon its DPMLR score, and is clearly one aspect that drives human detection performance. Future work is needed to identify a more meaningful FIQM. Such a measure may incorporate aspects of the contrast as well as other quality features of both the luminance and color components of the image. However, we expect that a measure needs to understand what context is available in the image, which makes the search for a good FIQM very challenging.

The paper revealed many interesting properties of the DPMLR and conjectured many more properties. Future work is necessary to prove (or disprove) these conjectured properties. Furthermore, one can further study over what values of $\tilde{\mathbf{p}}$ the DPMLRT is the most powerful test.

The DPMLRT does incorporate some simplifying assumptions that could be relaxed for a more robust test. For instance, not all human observers are created equal and the binomial distribution may not be the best model for the perception results. Furthermore, the values of $\tilde{p}_i$ are not independent since all fusion algorithms attempt to provide a good image for human perception. The paper does demonstrate that the DPMLRT is robust as these model assumptions are relaxed. In addition, the DPMLRT assumes that the observers' probability of false alarms are calibrated, and it ignores the impact of contextual information, which may be known *a priori* or obtained in the image, on human detection performance. Future research can also focus on statistical scoring mechanisms that account for increasingly realistic data models.

Appendix

A. Proof of Property 1

Proof: $\lambda_N^{\uparrow}(\mathbf{y},\mathbf{o})$ and $\lambda_N^{\downarrow}(\mathbf{y},\mathbf{o})$ can be expressed as

$$\lambda_N^{\uparrow}(\mathbf{y},\mathbf{o}) = \frac{N! \int_{\mathcal{P}_\uparrow} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p}}{\int_{\mathcal{P}_0} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p}} \tag{24}$$

and

$$\lambda_N^{\downarrow}(\mathbf{y},\mathbf{o}) = \frac{N! \int_{\mathcal{P}_\downarrow} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p}}{\int_{\mathcal{P}_0} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p}}. \tag{25}$$

Note that the integrands in the numerator and denominator are the same. This integrand is strictly positive for all $\mathbf{p}$ except for a finite set of points of measure zero, namely $\{\mathbf{p} : p_i \in \{0,1\}\}$. Any integral of the integrand over $\mathcal{P}_0$, $\mathcal{P}_\uparrow$, $\mathcal{P}_\downarrow$, $\mathcal{P}_0 \backslash \mathcal{P}_\uparrow$, and $\mathcal{P}_0 \backslash \mathcal{P}_\downarrow$ must be strictly positive. Thus, the integrals in the numerator of (24) are strictly less than the integrals in the denominator. Furthermore, all the integrals are strictly positive. Thus

$$0 < \frac{\int_{\mathcal{P}_\uparrow} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p}}{\int_{\mathcal{P}_0} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p}} < 1 \tag{26}$$

and

$$0 < \frac{\int_{\mathcal{P}_\downarrow} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p}}{\int_{\mathcal{P}_0} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p}} < 1. \tag{27}$$

Multiplication by $N!$ leads to $0 < \lambda_N^{\uparrow}(\mathbf{y},\mathbf{o}) < N!$ and $0 < \lambda_N^{\downarrow}(\mathbf{y},\mathbf{o}) < N!$. Because the ascending and descending DPMLRs are bounded by zero and $N!$ for each scene, it is clear by (11) that the composite DPMLR is also bounded by 0 and $N!$. ■

B. Proof of Property 3

Proof: Let $\pi_k : \{1,2,\ldots,N\} \longmapsto \{1,2,\ldots,N\}$ be a permutation mapping such that $\pi_k(i) \neq \pi_k(j)$ when $i \neq j$. There are $N!$ such mappings, and let each mapping be identified with a unique index k where $k = 1,2,\ldots,N!$. As a matter of convention, $k = 1$ is the identity mapping, i.e., $\pi_1(i) = i$, and $k = N!$ is the reverse sort, i.e., $\pi_{N!}(i) = N + 1 - i$. Each permutation function allows one to define an ordering of the coordinates, i.e.,

$$\mathcal{R}_k = \{\mathbf{p} : 0 \leqslant p_{\pi_k(1)} \leqslant p_{\pi_k(2)} \leqslant \ldots \leqslant p_{\pi_k(N)} \leqslant 1\} \tag{28}$$

such that the collection of all $N!$ orderings defines any possible sequence of coordinate values, i.e.,

$$\mathcal{P}_0 = \bigcup_{k=1}^{N!} \mathcal{R}_k. \tag{29}$$

Furthermore, $\mathcal{P}_\uparrow = \mathcal{R}_1$ and $\mathcal{P}_\downarrow = \mathcal{R}_{N!}$. As a result

$$\int_{\mathcal{P}_0} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p} = \sum_{k=1}^{N!} \int_{\mathcal{R}_k} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p}. \quad (30)$$

Using the change of variable $p_{\pi_k(i)} \mapsto p_i$, the right hand side of (30) can be rewritten as

$$\int_{\mathcal{P}_0} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p} = \sum_{k=1}^{N!} \int_{\mathcal{P}_\uparrow} \prod_{i=1}^{N} p_i^{y_{\pi_k(i)}}(1-p_i)^{o-y_{\pi_k(i)}} d\mathbf{p}. \quad (31)$$

If $y_1 = y_2 = \cdots = y_N$, then we have $y_{\pi_1(i)} = y_{\pi_2(i)} = \ldots = y_{\pi_{N!}(i)} = y_i$ for $i = 1, 2, \ldots, N$. It follows that:

$$\int_{\mathcal{P}_0} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p} = N! \int_{\mathcal{P}_\uparrow} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p}. \quad (32)$$

According to (24), we have $\lambda_N^{\downarrow}(\mathbf{y}, o\mathbf{1}) = 1 = \lambda_N^{\uparrow}(\mathbf{y}, o\mathbf{1})$ ■

C. *Proof of Properties 4 and 5*

Proof: First we want to show that if the y_i's are in ascending order and not constant, then

$$\prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} > \prod_{i=1}^{N} p_i^{y_{\pi_k(i)}}(1-p_i)^{o-y_{\pi_k(i)}} \quad (33)$$

when $k \neq 0$.

To this end, we transform the permutation π_k back to the identity. Let's define the permutation function $g_0(i) = \pi_k(i)$. For the first step, the value of $g_0(1)$ is switched with the value $g_0(j)$ where $g_0(j) = 1$ to form g_1. The process repeats itself for $N-1$ steps such that for the nth step, the value of $g_{n-1}(n)$ is switched with $g_{n-1}(j)$ where $g_{n-1}(j) = n$ to form g_n. Formally, at the nth step we have $g_n(n) = n, g_n(j) = g_{n-1}(n)$, and $g_n(i) = g_{n-1}(i)$, where $j = g_{n-1}^{-1}(n)$, $i > n$ and $i \neq j$.

Note that $j \geqslant n$ and $g_{n-1}(n) \geqslant n$ because $g_{N-1}(i) = i$ for $i = 1, 2, \ldots, n-1$. After $N-1$ steps, $\pi_1 = g_{n-1}$. After the nth step, the ratio of the likelihoods associated with permutations g_n and g_{n-1}, i.e.,

$$\frac{\prod_{i=1}^{N} p_i^{y_{g_n(i)}}(1-p_i)^{o-y_{g_n(i)}}}{\prod_{i=1}^{N} p_i^{y_{g_{n-1}(i)}}(1-p_i)^{o-y_{g_{n-1}(i)}}} = \left(\frac{p_n(1-p_j)}{p_j(1-p_n)}\right)^{y_n - y_{g_{n-1}(n)}} \quad (34)$$

is greater than or equal to unity because $y_n \leqslant y_{g_{n-1}(n)}$ and $p_n \leqslant p_j$ over $\mathcal{P}_\uparrow$. By taking the product of (34) for $n = 1, 2, \ldots, N-1$, we have

$$\frac{\prod_{i=1}^{N} p_i^{y_{g_{N-1}(i)}}(1-p_i)^{o-y_{g_{N-1}(i)}}}{\prod_{i=1}^{N} p_i^{y_{g_0(i)}}(1-p_i)^{o-y_{g_0(i)}}} \geqslant 1. \quad (35)$$

The equality occurs only if $y_n = y_{g_{n-1}(n)}$ for $n = 1, \ldots, N$, which means y_i's are equal. Because $g_0 = \pi_k$ and g_{N-1} is the identity map and the y_i's are not constant, (33) is proven.

Now, integrating both sides of (33) over $\mathcal{P}_\uparrow$ leads to

$$\int_{\mathcal{P}_\uparrow} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p} > \int_{\mathcal{P}_\uparrow} \prod_{i=1}^{N} p_i^{y_{\pi_k(i)}}(1-p_i)^{o-y_{\pi_k(i)}} d\mathbf{p} \quad (36)$$

when $k \neq 1$. Similarly, one can show that

$$\int_{\mathcal{P}_\uparrow} \prod_{i=1}^{N} p_i^{y_{\pi_{N!}(i)}}(1-p_i)^{o-y_{\pi_{N!}(i)}} d\mathbf{p} < \int_{\mathcal{P}_\uparrow} \prod_{i=1}^{N} p_i^{y_{\pi_k(i)}}(1-p_i)^{o-y_{\pi_k(i)}} d\mathbf{p} \quad (37)$$

when $k \neq N!$. The division of (36) and (37) by $\prod_{i=1}^{N} \beta(y_i + 1, o - y_i + 1)$ leads to the first statement in Property 4. Similar arguments prove the second statement in Property 4.

Summing (36) for $k = 1, \ldots, N!$ leads to

$$\int_{\mathcal{P}_0} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p} < N! \int_{\mathcal{P}_\uparrow} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p}. \quad (38)$$

Then $\lambda_N^{\uparrow}(\mathbf{y}, \mathbf{o}) > 1$. Similarly, (37) can be reexpressed as

$$\int_{\mathcal{P}_\downarrow} \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{o-y_i} d\mathbf{p} < \int_{\mathcal{P}_\downarrow} \prod_{i=1}^{N} p_i^{y_{\pi_k(i)}}(1-p_i)^{o-y_{\pi_k(i)}} d\mathbf{p} \quad (39)$$

so that $\lambda_N^{\downarrow}(\mathbf{y}, \mathbf{o}) < 1$. This completes the proof of the first statement in Property 5. The proof of the second statement can be proven by similar arguments. ■

ACKNOWLEDGMENT

The authors would like to thank R. K. Moore, University of Memphis. for registering and fusing the imagery. They would also like to thank Q. Nguyen, NVESD, for conducting the perception experiment.

REFERENCES

[1] Z. Zhang and R. S. Blum, "A region-based image fusion scheme for concealed weapon detection," in *Proc. 31st Annu Conf. Inf. Sci. Syst.*, 1997, pp. 168–173.

[2] G. Simone, A. Farina, F. C. Morabito, S. B. Serpico, and L. Bruzzone, "Image fusion techniques for remote sensing applications," *Inf. Fusion*, vol. 3, pp. 3–15, 2002.

[3] J. A. Castellanos, J. Neira, and J. D. Tardos, "Multisensor fusion for simultaneous localization and map building," *IEEE Trans. Robot. Autom.*, vol. 17, no. 6, pp. 908–914, Dec. 2001.

[4] S. P. Constantinos, M. S. Pattichis, and E. Mitheli-Tzanakou, "Medical imaging fusion applications: An overview," in *Proc. 35th Asilomar Conf. Signals Syst. Comput.*, 2001, vol. 2, pp. 1263–1267.

[5] R. R. Murphy, "Sensor and information fusion improved vision-based vehicleguidance," *IEEE Intell. Syst.*, vol. 13, no. 6, pp. 49–56, Nov./Dec. 1998.

[6] R. S. Blum and Z. Liu, *Multi-Sensor Image Fusion and Its Applications*. Boca Raton, FL: CRC Press, 2006.

[7] R. S. Blum, "On multisensor image fusion performance limits from an estimation theory perspective," *Inf. Fusion*, vol. 7, no. 3, pp. 250–263, 2006.

[8] Y. Chen, Z. Xue, and R. S. Blum, "Theoretical analysis of an information-based quality measure for image fusion," *Inf. Fusion*, vol. 9, no. 2, pp. 161–175, 2008.

[9] N. Cvejic, A. Loza, D. Bull, and N. Canagarajah, "A similarity metric for assessment of image fusion algorithms," *Int. J. Signal Process.*, vol. 2, no. 3, pp. 178–182, 2005.

[10] V. Petrovic and C. Xydeas, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, pp. 308–309, 2000.

[11] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electron. Lett.*, vol. 38, no. 7, pp. 313–315, Mar. 2002.

[12] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proc. Int. Conf. Image Process.*, 2003, vol. III, pp. 173–176.

[13] A. Toet, N. Schoumans, and J. K. Uspeert, "Perceptual evaluation of different nighttime imaging modalities," in *Proc. 3rd Int. Conf. Inf. Fusion*, 2000, vol. 1, pp. TUD3/17–TUD3/23.

[14] C. Wei and R. S. Blum, "Theoretical analysis of correlation-based quality measures for weighted averaging image fusion," *Inf. Fusion*, to be published.

[15] J. Puzicha, J. M. Buhmann, Y. Rubner, and C. Tomasi, "Empirical evaluation of dissimilarity measures for color and texture," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, 1999, vol. 2, pp. 1165–1173.

[16] Y. Chen and R. S. Blum, "A new automated quality assessment algorithm for image fusion," *Image Vis. Comput.*, vol. 27, no. 10, pp. 1421–1432, Sep. 2009.

[17] L. M. Kaplan, R. S. Blum, and S. D. Burks, "Analysis of image quality for image fusion via monotonic correlation," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 222–235, Apr. 2009.

[18] F. Sadjadi, "Comparative image fusion analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, Jun. 2005, vol. 3, pp. 8–15.

[19] D. Schmieder and M. Weathersby, "Detection performance in clutter with variable resolution," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-19, no. 4, pp. 622–630, Jul. 1983.

[20] L. M. Kaplan, "Extended fractal analysis for texture classification and segmentation," *IEEE Trans. Image Process.*, vol. 8, no. 11, pp. 1572–1585, Nov. 1999.

[21] M. J. T. Smith and A. Docef, *A Study Guide for Digital Image Processing*. Atlanta: GA Scientific Publishers, 1997.

[22] C. Howell, R. Moore, S. Burks, and C. Halford, "An evaluation of fusion algorithms using image fusion metrics and human identification performance," in *Proc. Infrared Imag. Syst. Design, Anal., Model., Test. XVIII*, G. C. Holst, Ed., 2007, vol. 6543, p. 65430V.

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, S. Member, and E. P. Simoncelli, "Image quality assessment: From error measurement to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[24] V. Laparra, J. Munoz-Mari, and J. Malo, "Divisive normalization image quality metric revisited," *J. Opt. Soc. Amer. A*, vol. 27, no. 4, pp. 852–864.

[25] W. Mendenhall, R. L. Scheaffer, and D. D. Wackerly, *Mathematical Statistics With Applications*, 3rd ed. Boston, MA: Duxbury Press, 1986.

[26] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*. New York: Halner Press, 1961.

[27] "Final report from the video quality experts group on validation of objective models of video quality assessment," Mar. 2000 [Online]. Available: http://www.vqeg.org/

[28] C. Wei, L. M. Kaplan, S. D. Burks, and R. S. Blum, "Diffuse prior monotonic likelihood ratio test for evaluation of fused image quality metrics," in *Proc. 12th Int. Conf. Inf. Fusion*, Seattle, WA, Jul. 2009, pp. 1076–1083.

[29] R. Barlow, D. Barholomew, J. Bremner, and H. Brunk, *Statistical Infernece Under Order Restrictions*. Hoboken, NJ: Wiley, 1972.

[30] M. Best and N. Chakravarti, "Active set algorithms for isotonic regression: A unifying approach," *Math. Program.*, vol. 47, pp. 425–439, May 1990.

[31] T. Moon and W. C. Stirling, *Mathematical Methods and Algorithms for Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 2000.

[32] E. Schumann, "Generating correlated uniform variates," 2009 [Online]. Available: http://comisef.wikidot.com/tutorial:correlateduniformvariates

[33] K. L. Kowalskia, "Generalized binomial distributions," *J. Math. Phys.*, vol. 41, no. 4, pp. 2375–2382, 2000.

[34] A. Toet, "Image fusion by a ratio of low-pass pyramid," *Pattern Recognit. Lett.*, vol. 9, no. 4, pp. 245–253, 1989.

[35] T. Huntsberger and B. Jawerth, "Wavelet based sensor fusion," in *Proc. SPIE*, 1993, vol. 2059, pp. 488–498.

[36] C. Lejeune, "Wavelet transform for infrared application," in *Proc. SPIE*, 1995, vol. 2552, pp. 313–324.

[37] L. Jiang, F. Tian, L. E. Shen, S. Wu, S. Yao, Z. Lu, and L. Xu, "Perceptual-based fusion of ir and visual images for human detection," in *Proc. Int. Symp. Intell. Multimedia, Video Speech Process.*, 2004, pp. 514–517.

[38] H. Chen and P. K. Varshney, "A human perception inspired quality metric for image fusion based on regional information," *Image Fusion*, vol. 8, no. 2, pp. 193–207, Apr. 2007.

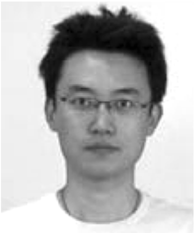
[39] O. O. Fadiran, P. Molnar, and L. M. Kaplan, "A statistical approach to quantifying clutter in hyperspectral infrared images," presented at the Proc. IEEE Aerosp. Conf., Big Sky, MT, Mar. 2006.

[40] D.-O. Kim and R.-H. Park, "New image quality metric using the Harris response," *IEEE Signal Process. Lett.*, vol. 16, no. 7, pp. 616–619, Jul. 2009.

[41] J. C. Leachtenauer, W. Malila, J. Irvine, L. Colburn, and N. Salvaggio, "General image-quality equation: GIQE," *Appl. Opt.*, vol. 36, no. 32, pp. 8322–8328, 1997.

[42] J. P. Estrera, "Localized signal-to-noise ratio of man and vehicle size targets," in *Proc. Infrared Technol. Appl. XXXV*, Orlando, FL, Apr. 2009, vol. 7298, p. 72 983M-72 983M-14.

[43] , J. Schanda, Ed., *Colorimetry: Understanding the CIE System*. Hoboken, NJ: Wiley, 2007.



Chuanming Wei received the B.Eng. and M.Eng. degrees in electrical engineering from University of Science and Technology of China (USTC), Hefei, China, in 2003 and 2006, respectively, and is currently pursuing the Ph.D. degree in electrical engineering at Lehigh University, Bethlehem, PA.

From 2003 to 2007, he worked as a research student at UT Starcom, Inc. His research interests include signal and image processing, multisensor data fusion, pattern recognition, radar, and sensor networking.

Lance M. Kaplan (S'88–M'89–SM'00) received the B.S. degree in electrical engineering (with distinction) from Duke University, Durham, NC, in 1989 and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, in 1991 and 1994.

From 1987 to 1990, he worked as a Technical Assistant at the Georgia Tech Research Institute, Atlanta, GA. He held a National Science Foundation Graduate Fellowship and a USC Dean's Merit Fellowship from 1990 to 1993, and worked as a Research Assistant in the Signal and Image Processing Institute at the University of Southern California from 1993 to 1994. He worked in the Reconnaissance Systems Department, Hughes Aircraft Company from 1994 to 1996. From 1996 to 2004, he was a member of the faculty in the Department of Engineering and a Senior Investigator in the Center of Theoretical Studies of Physical Systems (CTSPS), Clark Atlanta University (CAU), Atlanta, GA. Currently, he is a Team Leader in the Networked Sensing and Fusion branch of the U.S. Army Research Laboratory, Adelphi, MD. His current research interests include signal and image processing, automatic target recognition, information/data fusion, and resource management.

Dr. Kaplan serves as Associate Editor-In-Chief and EO/IR Systems Editor for the IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS. In addition, he is the tutorials editor for the *IEEE Aerospace and Electronic Systems Magazine*, and he also serves on the Board of Governors of the IEEE AES Society. He is a three time recipient of the Clark Atlanta University Electrical Engineering Instructional Excellence Award from 1999 to 2001.

Stephen D. Burks received the B.S. degree in physics from Emory University, Atlanta, GA, in 2000 and the M.S. degrees in physics, electrical and computer engineering, and public policy, respectively, from the Georgia Institute of Technology, Atlanta, GA, in 2002, 2002, and 2004, respectively.

He is currently the Team Leader of the Advanced Sensor Evaluation Facility (ASEF) at the U.S. Army Night Vision and Electronic Sensors Directorate, Fort Belvoir, VA. During his time as the leader of the ASEF, he has measured and characterized the performance of over 400 mid-wave and long-wave thermal imaging systems. His research interests include novel thermal imaging systems, image fusion, and thermal sensor performance model development.

Rick S. Blum (S'83–M'84–SM'94–F'05) received the B.S. degree in electrical engineering from the Pennsylvania State University, University Park, PA, in 1984 and the M.S. and Ph.D. degrees in electrical engineering from the University of Pennsylvania, Philadelphia, in 1987 and 1991, respectively.

From 1984 to 1991, he was a Member of technical staff of General Electric Aerospace, Valley Forge, PA, and graduated from GE's Advanced Course in Engineering. Since 1991, he has been with the Electrical and Computer Engineering Department, Lehigh University, Bethlehem, PA, where he is currently a Professor and holds the Robert W. Wieseman Chaired Research Professorship in Electrical Engineering. His research interests include signal processing for communications, sensor networking, radar and sensor processing.

Dr. Blum is on the editorial board for the *Journal of Advances in Information Fusion*. He was an associate editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and for IEEE COMMUNICATIONS LETTERS. He has edited special issues for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. He is a member of the SAM Technical Committee (TC) of the IEEE Signal Processing Society. He was a member of the Signal Processing for Communications TC of the IEEE Signal Processing Society and is a member of the Communications Theory TC of the IEEE Communications Society. He was on the awards Committee of the IEEE Communications Society. He is an IEEE Third Millennium Medal winner, a member of Eta Kappa Nu and Sigma Xi, and holds several patents. He was awarded an ONR Young Investigator Award in 1997 and an NSF Research Initiation Award in 1992. His IEEE Fellow Citation "for scientific contributions to detection, data fusion and signal processing with multiple sensors" acknowledges some early contributions to the field of sensor networking.